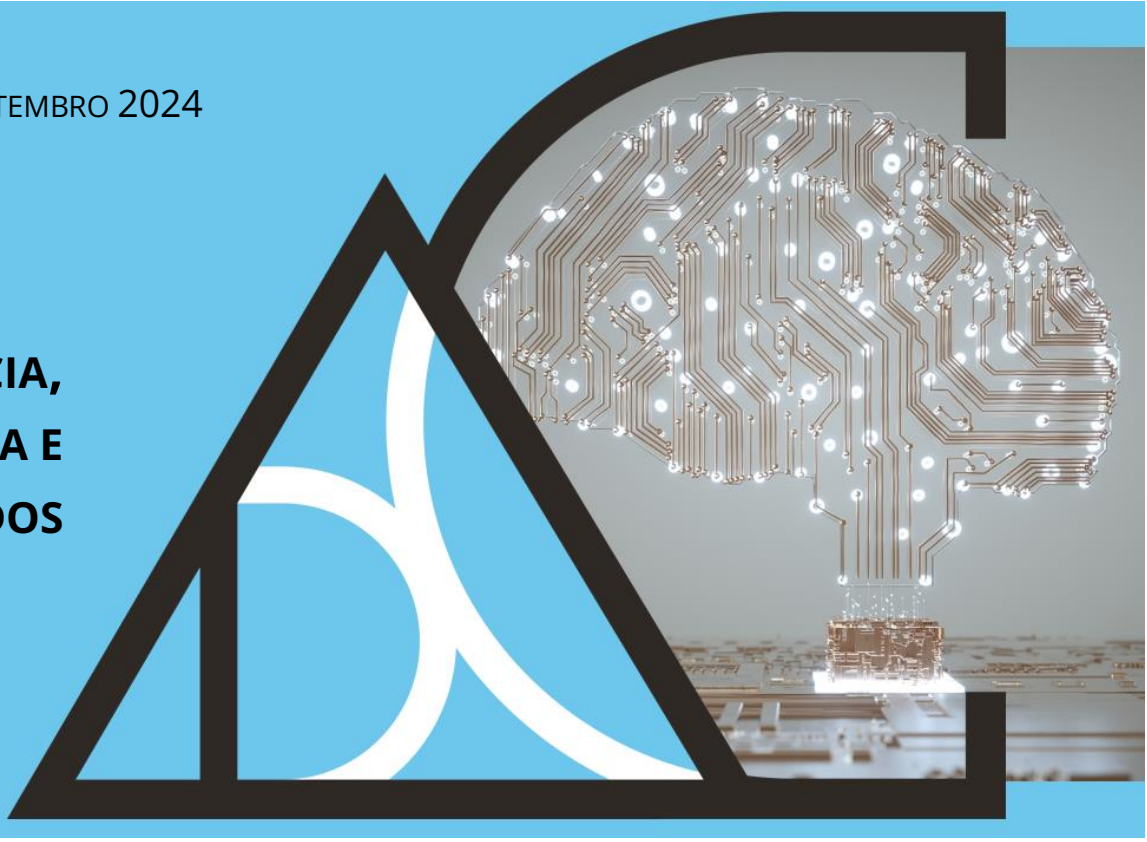
ADC SHORT PAPERS | SETEMBRO 2024

CONCORRÊNCIA, IA GENERATIVA E DADOS



Desde finais de 2022, a inteligência artificial (IA) generativa tem vindo a redefinir o setor digital. Trata-se de um novo tipo de IA capaz de produzir conteúdos semelhantes aos que um humano faria.

Em novembro de 2023, a Autoridade da Concorrência (AdC) publicou um *issues paper*[1] para acompanhar estes desenvolvimentos, mapeando os principais determinantes da concorrência na IA generativa e identificando os principais riscos para a concorrência no setor.

Este pequeno *paper* expande esse exercício, analisando a evolução do setor desde novembro de 2023. O documento incide, em particular, sobre a utilização de dados no desenvolvimento da IA generativa e a importância crescente de acordos de licenciamentos de dados.

I. Introdução

Os dados são um requisito essencial para o desenvolvimento de IA generativa, conjuntamente com a capacidade de computação e o *know-how*. Os dados apresentam-se em vários formatos (texto, imagem, vídeo, áudio, etc.), dependendo do tipo de modelo de IA generativa desenvolvido. Durante o treino, os padrões nos dados são incorporados no modelo de IA generativa, permitindo-lhe gerar novos conteúdos ao replicar esses padrões**.** Estes **dados de treino** podem ter diversas fontes e, como se explica *infra*, podem implicar custos de aquisição. Ademais, o conhecimento embutido nos modelos pode ser combinado com fontes de informação externas e verificáveis (*grounding*), tais como resultados de pesquisa, para melhorar a sua fiabilidade, a abrangência do conhecimento embutido no modelo e o acesso a informação mais atualizada. Desse modo, é possível reduzir a probabilidade de o modelo "alucinar".[2,3] Por último, os fornecedores de IA também recolhem **dados de monitorização** sobre o treino e o desempenho dos seus

[1] Disponível aqui.
[2] As alucinações, em grandes modelos de linguagem (Large Language Models – LLM), são respostas incorretas, enganadoras ou sem sentido de modelos de IA, mas que são apresentadas como factuais.
[3] No caso dos grandes modelos de linguagem, isto é feito através de técnicas de *retrieval-augmented generation* (RAG). Esta técnica é utilizada em serviços como o ChatGPT, o Perplexity AI ou o You.com. *Vide* ainda o artigo Lewis et al. (2021) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Disponível aqui.

modelos, por exemplo, observando o comportamento de utilizadores, para experimentar e otimizar futuras iterações dos modelos.[4]

Um *level playing field* no acesso a base de dados de grande dimensão, diversificadas, atualizadas e de elevada qualidade será crucial para promover a concorrência nos mercados de IA generativa.[5] Os modelos tipicamente têm um melhor desempenho em tarefas às quais já foram expostos durante o treino. Nessa medida, a riqueza dos dados de treino pode afetar de forma significativa o desempenho dos modelos,[6] especialmente no que diz respeito a pedidos de *long-tail*.[7] Estes são pedidos específicos que, individualmente, muito poucas pessoas fazem, mas que, coletivamente, representam uma grande parte do total dos pedidos. Neste tipo de pedidos, é mais provável que os modelos divirjam e produzam erros de geração, como alucinações em grandes modelos de linguagem (Large Language Models – LLM).[8]

A prossecução deste objetivo exige que as autoridades da concorrência identifiquem possíveis estrangulamentos no acesso aos dados no desenvolvimento da IA generativa. É importante determinar se os dados de treino não são substituíveis ou são difíceis de reproduzir pelos concorrentes, de modo que um pequeno número de empresas possa ser capaz de criar e explorar estrangulamentos em mercados de IA, de forma a prejudicar a concorrência, a inovação e, em última instância, os consumidores. Por outro lado, se for possível extrair informação semelhante de diferentes fontes de dados, o risco de estrangulamento será menor.

A identificação de eventuais estrangulamentos deve ter em conta as tendências recentes no que respeita à utilização de dados em IA generativa:

- **Até recentemente, os fornecedores de IA generativa utilizavam sobretudo dados públicos para treinar modelos de IA. Os fornecedores de IA têm vindo a tornar-se cada vez menos transparentes relativamente aos dados de treino que utilizam.**[9] Por este motivo, é mais difícil avaliar a existência de exclusividades no acesso aos dados, de acordos de licenciamento de dados e a importância de

[4] Para mais informação, *vide* as secções II, III e III.1 do Issues Paper da AdC, aqui.
[5] *Vide*, e.g., Longpre et al. (2023) A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity. Disponível aqui.
[6] *Vide*, e.g., Kandpal et al. (2023). Large Language Models Struggle to Learn Long-Tail Knowledge. Disponível aqui.
[7] O *long tail* refere-se a uma propriedade estatística em que um número grande de ocorrências aparece com pouca frequência, mas representa uma proporção significativa do total. Esta é, por exemplo, uma característica das vendas no comércio eletrónico. Existem muitos produtos que registam poucas vendas, mas correspondem a uma parcela significativa do total das vendas.
[8] Acontece um fenómeno semelhante em motores de busca, em que a escala de dados pode aumentar a qualidade de pesquisa, especialmente em buscas *long-tail* ou páginas menos visitadas. *Vide*, e.g., a discussão no Anexo I do estudo da CMA "*Online platforms and digital advertising market study*", disponível aqui.
[9] Por exemplo, para o GPT-3, o OpenAI indica os nomes das bases de dados que utilizou. Para o GPT-4, o OpenAI apenas refere que utiliza dados publicamente disponíveis e dados licenciados junto de terceiros. Por sua vez, para o GPT-4o, o OpenAI não fornece detalhes. A mesma evolução aconteceu entre o Llama 2 e o Llama 3, em que o Meta passou de uma breve descrição de cada base de dados para apenas mencionar que os dados de treino são obtidos a partir de fontes publicamente disponíveis. Esta evolução é também sinalizada em Longpre et al. (2023). The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. Disponível aqui.

cada base de dados para o desenvolvimento de modelos com bom desempenho.[10]

- **Os acordos de licenciamento de dados parecem ter-se tornado mais frequentes.** Tratam-se de acordos entre detentores de dados – como *publishers*, repositórios de imagens *stock* ou redes sociais – e fornecedores de IA generativa.
- **Os dados sintéticos e o pré-processamento de dados parecem estar a desempenhar um papel cada vez mais importante** no treino de modelos de IA generativos eficientes e com bom desempenho.[11]

Os capítulos seguintes tomam estas tendências como ponto de partida e identificam os determinantes de concorrência relacionados com a utilização de dados em IA generativa.

II. Questões de propriedade intelectual dos dados

Uma parte significativa dos dados de treino para modelos de IA generativa está publicamente disponível. Isto inclui páginas, imagens ou vídeos extraídos da Internet, bem como repositórios de livros. Na medida em que os dados de treino estão publicamente disponíveis, o acesso aos mesmos é mais equitativo e as principais barreiras são os recursos computacionais e os conhecimentos necessários para trabalhar com os dados.

No entanto, os dados publicamente disponíveis usados pelos fornecedores de IA generativa podem estar sujeitos a direitos de propriedade intelectual (PI). À medida que os modelos de negócio e as aplicações comerciais se têm consolidado e amadurecido, os detentores de PI começaram a exigir compensação pela utilização dos seus conteúdos. Os titulares de PI argumentam que os fornecedores de IA utilizam os seus conteúdos sem autorização durante o treino e a inferência dos modelos de IA, e que os sistemas de IA podem reproduzir ou gerar conteúdos derivados com base na sua PI. Como tal, alguns titulares de PI começaram a implementar ferramentas que protegem os seus conteúdos contra a utilização

[10] O Regulamento AI (AI Act) poderá conter algumas disposições relevantes a este respeito. Por exemplo, nos termos do regulamento, os fornecedores de um sistema de IA de alto risco devem fornecer documentação técnica do modelo, incluindo uma descrição das bases de dados de treino utilizadas, informação sobre a sua proveniência, âmbito e características principais; sobre como os dados foram obtidos e selecionados; sobre os procedimentos de rotulagem e as metodologias de limpeza dos dados (ver, por exemplo, o anexo IV referido no n.º 1 do artigo 11 do AI Act). Esta disposição, no âmbito do Capítulo 3, entra em vigor em 2 de agosto de 2025 (artigo 113.º).

[11] *Vide* secções II e III do presente documento.

não autorizada.[12]

Existe, por isso, um risco jurídico para os fornecedores de IA associado à utilização de uma grande parte dos dados de treino publicamente disponíveis. Permanece em aberto saber se os fornecedores de IA só podem utilizar dados abrangidos por direitos de PI se forem autorizados pelos titulares de PI. Tal dependerá da legislação existente em matéria de PI e da forma como esta é interpretada. Já foram intentadas inúmeras ações judiciais por titulares de PI contra fornecedores de IA generativa por violação de direitos de autor.[13]

Para atenuar o risco jurídico de violação de direitos de autor, os fornecedores de IA generativa têm vindo a celebrar acordos de licenciamento de dados com os titulares de PI. No que diz respeito ao setor da IA, os titulares de PI são produtores e/ou distribuidores de dados. Os titulares de PI publicam e distribuem frequentemente volumes significativos de conteúdos originais que são fundamentais para o treino de modelos de IA generativa (ver Caixa 1).

Caixa 1 – Exemplos de conteúdo licenciado

Os ***publishers*** **de notícias** produzem grandes volumes de texto relativamente formal, que pode ser fundamental para ensinar aos LLM a estrutura das diferentes línguas e factos sobre o mundo. Por exemplo, a Alex Springer, a Associated Press, o Financial Times, o Le Monde, o News Corp e a Prisa Media celebraram acordos de licenciamento de dados com a OpenAI.

Os LLM podem aprender linguagem informal e alargar os temas a que estão expostos através das **redes sociais**, onde os utilizadores publicam texto sob a forma de *posts* e comentários. Por exemplo, o Reddit celebrou acordos de licenciamento de dados com a Google e a OpenAI, e há notícias de conversações com a Meta e a Apple.

Do mesmo modo, o acesso a **repositórios de imagens *stock*** pode ser crucial para os modelos de geração de imagens, uma vez que estes produzem ou distribuem grandes volumes de imagens.

O mesmo se aplica às **plataformas de partilha de vídeos** e aos modelos de geração de vídeos. A Shutterstock, por exemplo, tem estado a licenciar as suas imagens, vídeos e música a criadores de IA como a OpenAI ou a Meta.

Tem havido uma vaga de acordos de licenciamento de dados. Isto sugere que os titulares de PI estão abertos a licenciar os seus dados a fornecedores de IA. Isto poderá proporcionar a alguns intervenientes no setor digital estratégias de monetização adicionais, como no caso de *publishers*, de redes sociais, de plataformas de partilha de vídeos ou de outras plataformas com muitos utilizadores que publicam conteúdos.

Já existem alguns exemplos públicos de acordos de dados, que têm sido avaliados em dezenas ou centenas de milhões de dólares.[14] Por exemplo, o Reddit revelou que está numa

[12] Por exemplo, os investigadores desenvolveram uma ferramenta que altera as imagens nos dados de treino. Quando utilizada, prejudica o desempenho do modelo de IA de geração de imagens. As alterações aplicadas às imagens não são percetíveis para um humano. *Vide* Shan et al. (2024). Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models. Disponível aqui.

[13] Por exemplo, New York Times v. OpenAI; Author's Guild v. OpenAI; a group of Artists v. Stability AI and v. Midjourney; ou Sony, Universal and Warner v. Suno and Udio.

[14] Por exemplo, a OpenAI pagará 250 milhões de dólares ao longo de cinco anos para aceder aos conteúdos da News Corp e a parceria entre a Google e o Reddit está avaliada em 60 milhões de dólares.

fase inicial da monetização dos seus dados e que estes serão fundamentais para o treino de LLM. Também referiu que celebrou acordos de licenciamento de dados no valor total de 203 milhões de dólares.[15] As plataformas podem também intensificar a recolha de dados. Por exemplo, em maio de 2024, a Meta anunciou que planeava utilizar publicações dos utilizadores do Facebook e do Instagram para treinar os seus futuros modelos de IA.[16] No entanto, isto não significa que os dados disponíveis nas plataformas digitais estejam necessariamente isentos de direitos de autor.[17]

Os titulares de PI podem preferir licenciar os seus dados para fins de *grounding* em vez de os licenciar apenas para o treino de modelos de IA. O *grounding* exige uma utilização recorrente dos seus dados, o que poderá assegurar um fluxo contínuo de receitas aos titulares de PI, ao passo que, durante o treino, os dados são normalmente utilizados poucas vezes.

A transição de dados publicamente disponíveis para dados proprietários e acordos de licenciamento de dados cria barreiras à entrada e à expansão, e pode reforçar o poder de mercado das empresas incumbentes com acesso a dados. Alguns fornecedores de IA podem dispor de melhores meios para adquirir dados e gerir os custos de transação associados ao licenciamento de dados. Adicionalmente, alguns titulares de PI podem optar por não vender os seus dados ou podem vendê-los de forma seletiva. Com efeito, alguns titulares de PI são também fornecedores de IA, devido à sua presença noutros mercados ou ecossistemas digitais. Por esta razão, podem ter menores incentivos para partilhar os seus dados, especialmente se os seus dados reforçarem ou sustentarem a sua posição em algum mercado digital.

As exclusividades no acesso aos dados de treino podem ser especialmente prejudiciais para a concorrência, sobretudo se os dados forem difíceis de substituir ou replicar. É provável que os modelos de IA generativa se diferenciem pela sua fiabilidade no *long-tail*. Isto sugere que o acesso a mais dados pode melhorar o desempenho do modelo, mesmo que existam neles redundâncias significativas, ou que a IA possa generalizar para além dos seus dados de treino. Os modelos de IA podem também diferenciar-se por se especializarem em determinadas tarefas, tópicos ou domínios. Se essas especializações exigirem dados específicos, as diferenças no acesso a esses dados, nomeadamente através de cláusulas de exclusividade ou de acesso discriminatório, podem criar barreiras à entrada.

Como tal, as exclusividades e o acesso preferencial a dados podem conferir aos fornecedores de IA vantagens concorrenciais indevidas, impedindo concorrentes de usar esses dados. Isto aumenta o poder de mercado e dificulta a inovação no setor da IA. Adicionalmente, estas exclusividades e acesso preferencial podem infringir a lei da concorrência em Portugal e na UE. Por exemplo, tal pode ocorrer se uma empresa tiver uma posição dominante num mercado de dados relevante e der acesso exclusivo ou preferencial a esses dados às suas próprias ofertas ou a terceiros, em detrimento dos concorrentes.

---

[15] *Vide* Reddit's Form S-1, submetido em fevereiro de 2024.
[16] *Vide* a publicação sobre esta alteração no blog da Meta, disponível aqui.
[17] Em junho, a Meta anunciou que tinha recebido um pedido da Comissão Irlandesa para a Proteção de Dados no sentido de adiar o treino de LLM utilizando conteúdos públicos partilhados por adultos no Facebook e no Instagram (*vide* aqui).

Um fluxo consistente de dados atualizados pode também ser importante para garantir modelos de IA com bom desempenho. Em algumas aplicações de IA, o valor dos dados pode diminuir com a idade, o que significa que são necessárias informações atualizadas para obter uma vantagem concorrencial face aos concorrentes. Tal pode aumentar o valor dos dados mais recentes e colocar os detentores de dados numa melhor posição para desenvolver modelos de IA que necessitem desses dados.

A promoção da concorrência e a garantia de um *level playing field* no acesso aos dados exigem processos agilizados de licenciamento de dados. Os acordos de dados bilaterais e personalizados podem aumentar os custos de transação e implicar barreiras significativas para os novos operadores no setor da IA. Ademais, as condições discriminatórias nos acordos de dados, como as exclusividades, podem exacerbar as barreiras à entrada e à expansão. Do mesmo modo, as estruturas de preços que exigem que os fornecedores de IA paguem antecipadamente pelos dados podem favorecer os operadores com mais capacidade financeira, especialmente no caso dos modelos-base.

Podem ser consideradas várias opções para agilizar o acesso aos dados. Por exemplo, opções como a disponibilização de dados através de API abertas, o agrupamento de licenças em pacotes e a adoção de estruturas de preços de pagamento em função da utilização (*pay-as-you-go*), para evitar efeitos de escala, podem ser formas eficazes de atenuar estas preocupações. Tornar os dados públicos facilmente disponíveis e sem restrições desnecessárias, como sejam os repositórios de livros de domínio público das bibliotecas nacionais ou decisões judiciais, pode também contribuir para reduzir as barreiras à entrada e à expansão da IA generativa.

A transição para dados proprietários pode reforçar a concentração

Uma vez que os titulares de PI passaram a exigir compensação, tem havido uma transição de dados publicamente disponíveis para dados proprietários, o que pode reforçar vantagens associadas a dados e a concentração no mercado.

A exclusividade dos dados pode ser prejudicial à concorrência

A exclusividade e o acesso preferencial a dados pode ser especialmente prejudicial para a concorrência e pode infringir a lei da concorrência em Portugal e na UE.

Agilizar o acesso a dados para garantir um *level playing field*

A agilização do acesso a dados a fornecedores de IA será crucial para garantir um *level playing field* (e.g., fornecendo dados via API abertas, estruturas de preços de *pay-as-you-go* ou facilitando o acesso a dados públicos).

III. Dados sintéticos

Dados sintéticos são dados gerados artificialmente por um algoritmo, nomeadamente modelos de IA generativa, que podem ser posteriormente usados para treinar novos modelos de IA generativa.[18]

[18] Os dados sintéticos também podem ser úteis para além do desenvolvimento de IA generativa. Por exemplo, programadores de modelos de *machine learning* podem recorrer a dados sintéticos se tiverem dificuldades de acesso a dados reais. Isto tem inúmeras

Conceptualmente, o uso de dados sintéticos segue um princípio semelhante ao *transfer learning* (e.g., *fine-tuning*)[19]. Com *transfer learning*, novos modelos são treinados a partir de outros modelos pré-treinados; com dados sintéticos o novo modelo é treinado usando dados gerados por um um modelo pré-treinado.

A utilização de dados sintéticos para ter-se disseminado, dado que muitos fornecedores de IA generativa recorrem a este de dados no desenvolvimento dos seus modelos. Por exemplo, a Meta usou dados sintéticos gerados pelo Llama 2 para treinar o modelo Llama 3.[20] De igual modo, a Anthropic usou dados sintéticos para treinar a família de modelos Claude 3.[21] A Nvidia também desenvolveu um modelo especializado em gerar dados sintéticos.[22] Por fim, a OpenAI criou um guia sobre como gerar dados sintéticos a partir do GPT-4.[23]

Os dados sintéticos podem ser uma alternativa a dados reais, efetivamente reduzindo barreiras à entrada, custos de aquisição de dados e constituindo uma alternativa para fornecedores de IA que não têm acesso a bases de dados específicas ou que querem aumentar a riqueza e diversidade dos dados de treino em certos domínios (*data augmentation*). Desse modo, o acesso a dados é mediado pelos modelos de IA generativa que geram dados sintéticos, e que replicam os padrões da base de dados original. Os dados sintéticos também podem permitir melhor proteger a privacidade, segredos de negócio ou outra informação sensível.[24] Adicionalmente, os dados sintéticos podem ser particularmente úteis dado serem mais estruturados e fáceis de usar no desenvolvimento de modelos de IA.

No entanto, depender em demasia de dados sintéticos pode diminuir o desempenho dos modelos, o que limita a eficácia dos dados sintéticos na redução de diferenças entre fornecedores de IA no acesso a dados. À medida que a proporção dos dados sintéticos nos dados de treino aumenta, o desempenho do modelo pode degradar-se. Isto pode acontecer pois os dados sintéticos têm um grau de distância da base de dados original, o que significa que erros de geração e enviesamentos do modelo original passam para o novo modelo enquanto dados de treino. Acresce que os dados sintéticos tendem a ser menos diversos que os dados reais subjacentes.[25] Para mitigar estas preocupações, os fornecedores de IA devem usar

---

aplicações, e.g. na área de saúde ou na deteção de fraude. Ver, e.g., Ktena et al. (2024). Generative models improve fairness of medical classifiers under distribution shifts; ou Benalcazar et al (2023). Synthetic ID Card Image Generation for Improving Presentation Attack Detection.

[19] *Vide* secção III do Issues Paper da AdC sobre IA Generativa, aqui.

[20] *Vide* a publicação de *blog* da Meta que introduz o Llama 3, aqui, e o artigo que introduz o Llama 3, aqui.

[21] *Vide* o artigo que introduz a família de modelos Claude 3, da Anthropic, aqui.

[22] *Vide* a publicação de blog da Nvidia que apresenta a família de modelos Nemotron-4 340B, aqui.

[23] *Vide* o artigo do OpenAI Cookbook, aqui.

[24] *Vide*, e.g., Afonja et al. (2024). The Crossroads of Innovation and Privacy: Private Synthetic Data for Generative AI. Disponível aqui.

[25] A literature denomina este problema de "colapso do modelo" (*model colapse*). Esta linha de investigação preocupa-se sobretudo com a possibilidade de os programadores usarem dados sintéticos sem querer nos seus dados de treino. Isto pode acontecer, porque os programadores recolhem dados da Internet que, por sua vez, podem ter sido gerados por IA. Ver um exemplo de colapso do modelo para modelos de linguagem (LLM) em Shumailov et al. (2024). The Curse of Recursion: Training on Generated Data Makes Models Forget; e Shumailov et al. (2024). AI models collapse when trained on recursively generated data. Disponíveis aqui e aqui. Este problema também foi identificado em modelos de IA generative de imagens – ver, e.g., Alemohammad et al. (2023). Self-Consuming Generative Models Go MAD. Disponível aqui.

dados sintéticos e dados reais em conjunto,[26] o que limita a capacidade de os dados sintéticos replicarem ou substituírem dados reais.

Os fornecedores de IA generativa podem introduzir limitações aos usos que outros fornecedores fazem de dados sintéticos. Por exemplo, nos seus termos de uso, a OpenAI não permite que os utilizadores usem as respostas dos seus modelos para competir com a OpenAI.[27] Adicionalmente, a licença do Llama 3 da Meta não permite a programadores usarem os dados gerados pelos modelos do Llama 3 para melhorar outros modelos de IA.[28] Estes termos podem limitar a viabilidade comercial ou a escala de modelos treinados com recurso a dados sintéticos. Em contraste, outros modelos, como o Nemotron-4 340B da Nvidia, são abertos e não introduzem restrições aos programadores.[29]

Os dados sintéticos irão provavelmente continuar a ser uma alternativa a dados reais, apesar das suas limitações. Os dados sintéticos podem, pelo menos parcialmente e em certa medida, substituir e replicar dados reais. Como tal, são capazes de mitigar algumas das diferenças no acesso a dados e criar um maior *level playing field* no setor. Ainda assim, não são suficientes para assegurar contestabilidade e os fornecedores de IA com acesso a dados reais podem ter uma vantagem competitiva sobre concorrentes.

Dados sintéticos não são garante de contestabilidade

Os dados sintéticos são cada vez mais usados por fornecedores de IA para reduzir barreiras à entrada e custos de aquisição de dados, mas têm limitações e os fornecedores com acesso a dados reais podem ter uma vantagem competitiva.

IV. Pré-processamento de dados

O pré-processamento de dados (em inglês, também denominado *data filtering* ou *data selection*) é um passo essencial no treino de qualquer modelo de IA e um fator diferenciador chave, dado que pode aumentar a qualidade dos modelos. O objetivo é curar dados em bruto, transformando-os em dados de maior qualidade mais adequado para o treino de IA, de modo a otimizar a eficiência e desempenho dos modelos.

Todos os principais modelos de IA generativa recorrem de algum modo a técnicas de pré-processamento de dados. Por exemplo, a Meta, no desenvolvimento do Llama 3, criou vários *pipelines* de filtragem de dados.[30] De igual modo, a Google aplica filtros nos dados de treino dos modelos Gemini[31] e a Microsoft otimizou a eficiência do modelo Phi-3 curando estrategicamente os dados de treino, entre outras técnicas.[32]

[26] *Vide*, e.g., Gerstgrasser et al. (2024). Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data. Disponível aqui.
[27] Europe Terms of Use, by OpenAI. Disponível aqui.
[28] *Vide* Meta Llama 3 Community License Agreement. Disponível aqui. Isto alterou-se com a publicação do Llama 3.1, cuja licença requer que os modelos desenvolvidos com recurso ao Llama 3.1 devem estar sob a mesma licença que o Llama 3.1, exibir a mensagem "Built with Llama" nos materiais de publicação do modelo e incluir "Llama" no início do nome do modelo. Ver Meta Llama 3.1 Community License Agreement. Disponível aqui.
[29] *Vide* nota de rodapé 23.
[30] *Vide* as publicações de *blog* da Meta que introduzem o Llama 3 e 3.1, aqui e aqui. *Vide* também o artigo que apresenta o Llama 3, aqui, onde a Meta detalha muitas da experimentação que fez durante o desenvolvimento do Llama 3.
[31] *Vide* o Technical Report do Gemini 1 e o Technical Report do Gemini 1.5, pela Google, aqui e aqui.
[32] *Vide* a publicação de *blog* da Microsoft que introduz o Phi-3, aqui.

Para maximizar o desempenho e a eficiência dos modelos de IA, os fornecedores de IA devem escolher um *mix* ótimo de técnicas de pré-processamento de dados. Existem diversas técnicas, incluindo remover dados de baixa qualidade, informação duplicada ou misturar dados de fontes diferentes de formas específicas (ver Caixa 2).

Caixa 2 – Exemplos de técnicas de pré-processamento de dados[33]

Os fornecedores de modelos LLM podem recorrer a um vasto conjunto de técnicas de pré-processamento de dados. Esse leque de opções é ilustrativo das muitas decisões que os fornecedores de IA têm de tomar, que podem depois ter impacto na eficiência e desempenho de modelos de IA generativa. Alguns exemplos:

- **Filtros de língua** (*language filtering*) selecionam documentos que apenas incluem as línguas pretendidas no modelo, como linguagens de programação.
- **Métodos heurísticos** podem ser úteis para remover grandes volumes de texto que não seja adequado para treinar modelos, como documentos com poucas palavras, linhas onde palavras se repetem muitas vezes, ou com muitos símbolos (e.g., # ou -) e poucas palavras.
- Os fornecedores de IA também podem filtrar de acordo com a **qualidade de dados**, escolhendo dados que são parecidos com outros que consideram ser de alta qualidade.
- Se os modelos forem treinados para um determinado domínio (e.g., medicina ou direito), os fornecedores de IA podem **filtrar por dados específicos a esse domínio**, comparando-os com bases de dados especializadas nesse domínio.
- **Remover dados duplicados** ou quase duplicados é um passo importante para melhorar a eficiência e desempenho dos modelos.
- **Filtrar conteúdo tóxico e explícito** remove conteúdo ilegal ou extremamente indesejável da base de dados de treino e induz os modelos a produzir menos conteúdo desse tipo.
- ***Mixing* de dados** atribui a cada base de dados maior ou menor importância, e pode ter um impacto significativo no desempenho dos modelos.

Otimizar o pré-processamento de dados requer esforços de experimentação significativos por parte dos fornecedores de IA. Isto exige recursos computacionais significativos, tempo e pessoal especializado.[34] Esta otimização é tipicamente baseada na intuição e no *know-how* dos programadores e das equipas, num processo de *learning-by-doing*. Estas técnicas são muitas vezes descritas como uma "arte" e são pouco documentadas.[35]

A experimentação é um fator estrutural do desenvolvimento de IA que multiplica os custos de desenvolvimento e os efeitos de

[33] Estes exemplos foram retirados de um inquérito às técnicas de pré-processamento de dados, em Albalak et al. (2024). A Survey on Data Selection for Language Models. Disponível aqui. Mais exemplos de técnicas de filtragem de dados estão disponíveis no artigo que apresenta o Llama 3 da Meta, aqui.
[34] *Vide* ainda secção III.1 e III.3 no Issues Paper da AdC sobre IA Generativa, aqui.
[35] *Vide*, e.g., a nota de rodapé 5.

escala dos modelos de IA generativa, tornando os mercados de IA mais propensos a concentração. Isto é especialmente verdade no caso de modelos-base, onde pode conduzir a cenários em que um pequeno número de modelos-base se torna um *input* crítico para mercados a jusante. Por esse motivo, a necessidade de experimentação no desenvolvimento de IA reforça a importância do acesso a *inputs* chave, como sejam a capacidade de computação, em reduzir barreiras à entrada.

Os modelos de IA em código aberto podem, neste contexto, desempenhar um papel chave em reduzir barreiras à entrada induzidas pela necessidade de experimentação. A questão-chave na experimentação é encontrar a combinação ótima de arquiteturas e hiperparâmetros dos modelos – e, no caso específico dos dados, o *mix* ótimo de técnicas de pré-processamento de dados. Uma vez encontradas, este conhecimento pode ser documentado e partilhado pela indústria. Devido à sua maior transparência, modelos em código aberto e documentação detalhada sobre os mesmos vão ser uma via importante para reduzir necessidades de experimentação e, desse modo, custos de desenvolvimento dos modelos.

Código aberto mitiga efeitos de escala devidos a experimentação

O pré-processamento de dados é crucial no treino de modelos de IA, mas pode reforçar a concentração. Canais de partilha de conhecimento, como modelos em código aberto, podem mitigar efeitos de escala devido a experimentação.

Concorrência, IA Generativa e Dados
Principais Mensagens