



ISSUES PAPER

**CONCORRÊNCIA E
INTELIGÊNCIA
ARTIFICIAL
GENERATIVA**

NOVEMBRO/2023

SUMÁRIO EXECUTIVO

A Inteligência Artificial Generativa (IA Generativa) é a inteligência artificial capaz de produzir conteúdos de forma similar a um ser humano, como sejam texto, imagem, vídeo, áudio e código, mas de forma massificada. Esta tecnologia capta padrões na informação utilizada para o seu treino e depois é capaz de os replicar. Isto confere à IA Generativa uma capacidade de generalização e de abstração tipicamente associada aos seres humanos.

A integração da IA Generativa nos processos produtivos e de inovação pode ter um impacto significativo na economia. Ao automatizar tarefas que antes exigiam uma extensa contribuição humana, esta tecnologia irá permitir reduzir o tempo de execução de várias tarefas, contribuindo positivamente para a produtividade. Estima-se que a IA Generativa possa acrescentar até 7% do PIB mundial em dez anos,¹ e que esteja presente na generalidade das empresas, como por exemplo no apoio ao cliente e na análise de dados, mas também em áreas como a educação, a investigação científica, a descoberta de medicamentos e o entretenimento.

A IA Generativa já é utilizada por milhões de utilizadores desde que vários serviços foram disponibilizados ao público, como seja o ChatGPT, Stable Diffusion, Copilot ou Midjourney. Estima-se que o serviço ChatGPT, da OpenAI, tenha atraído 1,5 mil milhões de visitas em todo o mundo apenas durante o mês de setembro de 2023.² Segundo um inquérito do Euroconsumers³ sobre a utilização do ChatGPT em quatro países, incluindo Portugal, 75% dos inquiridos com idades entre 18 e 34 anos, e 29% do grupo entre 55 e 74 anos, referiram já ter experimentado o ChatGPT. A principal razão para os inquiridos utilizarem este serviço ou sistemas baseados neste é a pesquisa de informações (68%), seguida da capacidade para gerar texto (62%), resumir textos longos (37%), obter inspiração (31%) ou gerar imagens (25%).

É possível expandir as funcionalidades da IA Generativa através de *plugins*, que podem ser equiparados a *apps* para IA. Os *plugins* podem ser desenvolvidos por terceiros e são semelhantes a aplicações para um sistema operativo. Por exemplo, um *plugin* onde um utilizador solicita à IA um roteiro de uma viagem, especificando data e local, e a IA responde com o roteiro, estadias, voos e um *link* para a agência de viagem onde fazer a reserva. Num outro *plugin*, um utilizador pode fazer perguntas sobre os conteúdos de um documento PDF por si submetido à IA, por exemplo, para resumir capítulos. O ChatGPT, por exemplo, já disponibiliza mais de 1000 *plugins*, tendo os primeiros sido desenvolvidos por Expedia,

¹ Goldman Sachs (2023), "[Generative AI could raise global GDP by 7%](#)", acessido a 17.10.2023.

² Estatísticas da similarweb, disponíveis [aqui](#), acessido a 31.10.2023.

³ Euroconsumers (2023) "[Me, myself and Generative AI](#)". O inquérito foi realizado online simultaneamente na Bélgica, Itália, Portugal e Espanha durante a primeira semana de junho de 2023, tendo sido recolhidas 4212 respostas válidas.

FiscalNote, Instacart, KAYAK, Klarna, Milo, OpenTable, Shopify, Slack, Speak, Wolfram e Zapier.⁴

A concorrência, enquanto promotor da inovação, é fundamental para concretizar as oportunidades criadas pela disrupção tecnológica associada à IA Generativa. Um ambiente concorrencial e contestável promove a inovação pelas empresas já estabelecidas no mercado e garante que operadores com produtos inovadores ou processos mais eficientes possam entrar e crescer, em benefício dos consumidores e da economia.

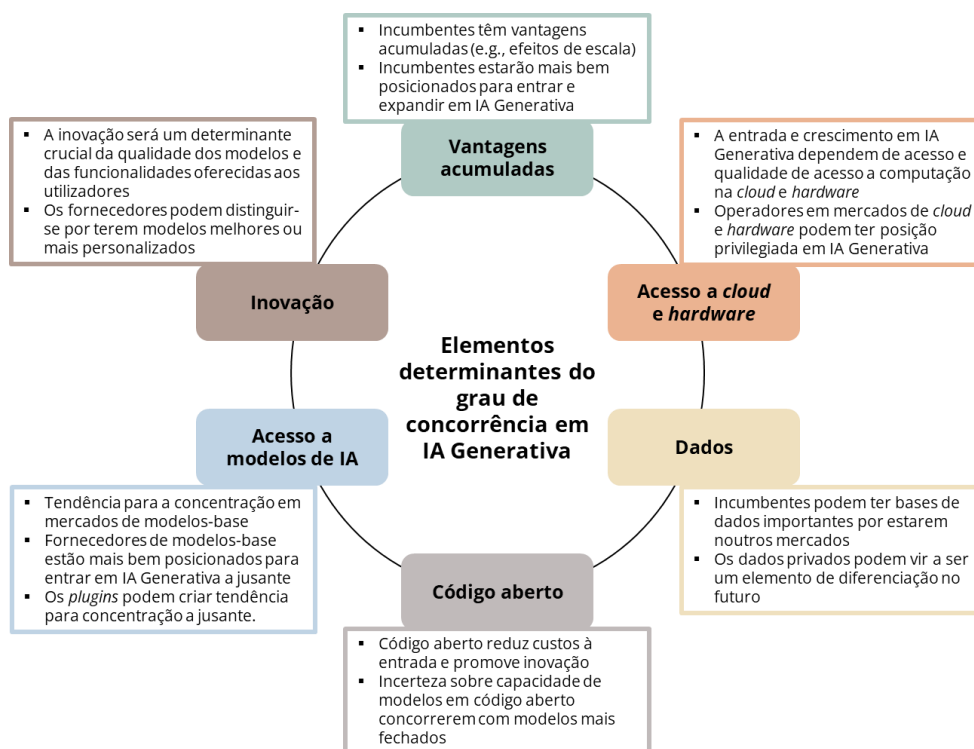
A importância de concretizar plenamente todos os benefícios da IA foi já reconhecida a nível da União Europeia com a proposta de Regulamento Inteligência Artificial, que visa tornar os sistemas de IA seguros, rastreáveis e respeitadores do ambiente. Nos Estados Unidos, este tema também se encontra na agenda de política pública. A 30 de outubro de 2023, o Presidente Biden emitiu uma *Executive Order* sobre a IA com vista a promover a equidade e os direitos civis, defender os consumidores e os trabalhadores, promover a inovação e a concorrência, e assegurar a segurança dos sistemas de IA.

Este surto de inovação e a corrida pela IA introduzem um momento de contestabilidade no mercado. O momento atual definirá a dinâmica concorrencial em IA Generativa no futuro, pelo que é importante assegurar um ambiente concorrencial desde o princípio. A título ilustrativo, no final de 2022, a OpenAI anunciou que o ChatGPT ultrapassou um milhão de utilizadores, em apenas cinco dias após ter sido disponibilizado ao público e estima-se que tenha ultrapassado 100 milhões de utilizadores após dois meses. Outros serviços digitais, e.g., Instagram ou Spotify, terão demorado pelo menos mais de dois meses e meio para atingir um milhão de utilizadores.⁵ Estas estimativas mostram a rápida adoção destas novas tecnologias e a importância de assegurar as condições de concorrência desde o princípio.

Neste Issues Paper, a AdC mapeia os elementos determinantes que afetam o processo concorrencial e antecipa os riscos para a concorrência, no setor da IA Generativa, baseando-se na experiência acumulada no setor digital.

⁴ OpenAI Blog (2023). "[ChatGPT plugins](#)", acedido a 22.10.2023.

⁵ De acordo com informação disponibilizada [aqui](#), [aqui](#) e [aqui](#).



A IA Generativa é um exemplo emblemático dos desafios de concorrência já identificados no setor digital. Os mercados de IA Generativa apresentam características que os tornam suscetíveis a elevados **níveis de concentração**. Os modelos são ávidos de dados e de computação, propiciando fortes efeitos de escala. Estes efeitos podem resultar em vantagens competitivas acumuladas para os incumbentes digitais, que já têm acesso a grandes volumes de dados e capacidade de computação. Tal como noutros mercados digitais, estas características podem trazer riscos de concorrência, nomeadamente em termos de estratégias de **exclusão de concorrentes**, em mercados de computação na *cloud*, de *hardware* e de modelos de IA Generativa. Os modelos de IA Generativa têm sido disponibilizados em formato de código aberto, mas poderão apresentar limitações quanto à sua capacidade para concorrer com modelos em formato mais fechado.

Como tal, o foco da defesa da concorrência no contexto da Inteligência Artificial terá necessariamente de estar nos seguintes elementos-chave: (i) acesso aos dados, (ii) acesso à computação na *cloud* ou *hardware* especializado e (iii) acesso aos modelos-base de IA Generativa.

A AdC, no âmbito das suas atribuições e no contexto da cooperação internacional, manter-se-á atenta aos desenvolvimentos da inteligência artificial, e não hesitará em intervir para garantir a promoção e a aplicação do direito da concorrência em Portugal, em benefício dos consumidores e sempre que os riscos identificados se materializem.

I. INTRODUÇÃO

Os recentes avanços na inteligência artificial (IA) Generativa – tecnologias que geram novos conteúdos, desde texto a código e imagens - têm sido considerados como um choque tecnológico com impactos notáveis na economia (cf. Caixa 1). Alguns dos seus contributos podem vir a passar pela produção de conteúdos de texto ou imagem a um custo inferior, identificação de anomalias e fraude, otimização de serviços de apoio ao cliente e aceleração de projetos de descoberta de medicamentos e terapêuticas.

A IA Generativa já é utilizada por milhões de utilizadores desde que vários serviços foram disponibilizados ao público, como seja o ChatGPT, Stable Diffusion, Copilot ou Midjourney (cf. Caixa 1).

Caixa 1 – Estatísticas quanto ao impacto e utilização da IA Generativa

Impacto estimado da IA Generativa na economia

- Embora ainda seja cedo, um relatório que a IA Generativa possa acrescentar até 7% do PIB mundial em dez anos.⁶
- Um outro relatório estima que, de um conjunto de 63 *use cases* (i.e., casos de aplicação), a IA generativa tem o potencial para gerar 2,6 biliões a 4,4 biliões de dólares em valor na economia. A banca, a tecnologia e a área de investigação de descoberta de medicamentos poderão registar o maior impacto em termos de percentagem das suas receitas da IA generativa. Por exemplo, o relatório estima que, no setor bancário, o recurso à IA Generativa poderá proporcionar um valor equivalente a mais 200 mil milhões a 340 mil milhões de dólares anuais se os *use cases* forem concretizados.⁷
- 40% dos executivos inquiridos, de um universo de 1000 organizações, identificaram já ter criado equipas e orçamentos dedicados à IA generativa, enquanto 49% está a pensar fazê-lo no prazo de 12 meses. A maioria dos executivos inquiridos identificaram as seguintes principais áreas onde IA Generativa terá maior impacto: (i) tecnologia (e.g., para gerar dados sintéticos); (ii) vendas e apoio ao cliente (e.g., via assistentes virtuais) e; (iii) marketing e comunicação (e.g., para efeitos de otimização de campanhas).⁸
- Um estudo empírico concluiu que o recurso a conversação baseada em IA aumentou a produtividade de um *call-center* em 14%, medida pelo número de questões de clientes que os agentes podem resolver por hora.⁹

⁶ Goldman Sachs (2023), "[Generative AI could raise global GDP by 7%](#)", acedido a 17.10.2023.

⁷ McKinsey & Company (2023) "[The economic potential of generative AI. The next productivity frontier](#)".

⁸ Capgemini Research (2023). "[Harnessing the value of generative AI: Top use cases across industries](#)". O inquérito foi realizado a executivos de 1000 organizações em abril de 2023 em formato online. Foram também realizadas entrevistas adicionais a 10 executivos.

⁹ Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). *Generative AI at work*.

Utilização de serviços de IA Generativa

- A IA Generativa já é utilizada por milhões de utilizadores desde que vários serviços foram disponibilizados ao público, como seja o ChatGPT, Stable Diffusion, Copilot ou Midjourney.
- No final de 2022, a OpenAI anunciou que o ChatGPT ultrapassou um milhão de utilizadores, em apenas cinco dias após ter sido disponibilizado ao público e estima-se que tenha ultrapassado 100 milhões de utilizadores após dois meses. Outros serviços digitais, e.g., Instagram ou Spotify, terão demorado pelo menos mais de dois meses e meio para atingir um milhão de utilizadores.¹⁰
- Estima-se que o serviço ChatGPT, da OpenAI, tenha atraído 1,5 mil milhões de visitas em todo o mundo apenas durante o mês de setembro de 2023.¹¹
- Segundo um inquérito do Euroconsumers¹² sobre a utilização do ChatGPT em quatro países, incluindo Portugal, 75% dos inquiridos com idades entre 18 e 34 anos, e 29% do grupo entre 55 e 74 anos, referiram já ter experimentado o ChatGPT. A principal razão para os inquiridos utilizarem este serviço ou sistemas baseados neste é a pesquisa de informações (68%), seguida da capacidade para gerar texto (62%), resumir textos longos (37%), obter inspiração (31%) ou gerar imagens (25%).

Existem, contudo, riscos de concorrência associados ao desenvolvimento e à implementação da IA Generativa. Tal como outros serviços digitais, os mercados de IA Generativa e alguns mercados adjacentes apresentam características que os tornam suscetíveis a elevados níveis de concentração. O desenvolvimento e a implementação de modelos-base são passíveis de estar sujeitos a fortes efeitos de escala.

Alguns operadores podem obter um poder de mercado significativo e ter a capacidade e o incentivo para introduzir e explorar estrangulamentos no mercado. Estes estrangulamentos baseiam-se, genericamente, no acesso e controlo de algum recurso necessário para colocar uma oferta competitiva no mercado. Exemplos incluem bases de dados, a capacidade de recolha de dados, a capacidade de computação, bases de utilizadores e efeitos de rede, ou a experiência acumulada do operador.

O presente Issues Paper aborda questões de concorrência que se colocam no âmbito da IA Generativa. Em particular, desenvolve-se uma análise sobre os requisitos necessários para o desenvolvimento e a implementação de serviços de IA Generativa. Aborda-se ainda a concorrência entre fornecedores de modelos-base e a concorrência a jusante. Identificam-se riscos de concorrência associados ao desenvolvimento e implementação da IA Generativa. Por fim, desenvolve-se um mapeamento de alguns dos determinantes do grau de concorrência no setor de IA Generativa.

¹⁰ De acordo com informação disponibilizada [aqui](#), [aqui](#) e [aqui](#).

¹¹ Estatísticas da similarweb, disponíveis [aqui](#), acedido a 31.10.2023.

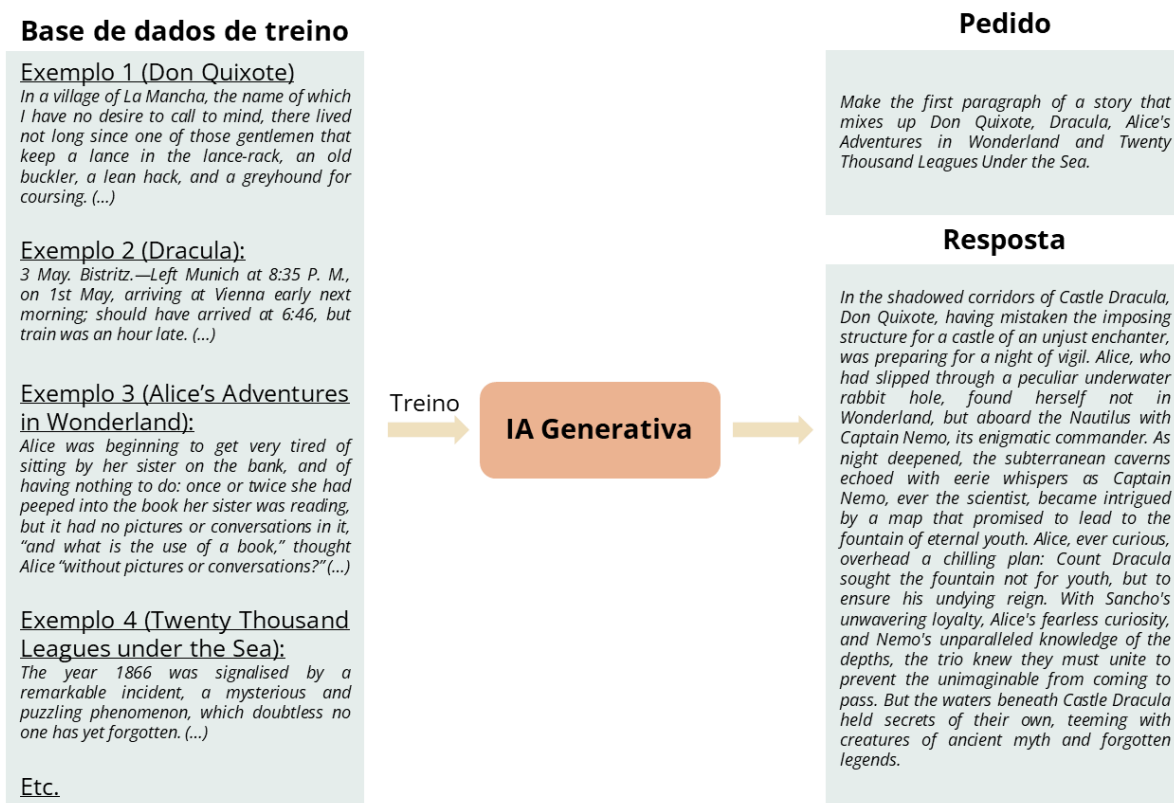
¹² Euroconsumers (2023) "[Me, myself and Generative AI](#)". O inquérito foi realizado online simultaneamente na Bélgica, Itália, Portugal e Espanha durante a primeira semana de junho de 2023, tendo sido recolhidas 4212 respostas válidas.

Como nota final, importa ressaltar que (i) os avanços na IA Generativa são recentes e só agora se começam a materializar os modelos de negócio e formas de monetização; e (ii) dadas as atribuições da Autoridade da Concorrência (AdC), não se analisam questões fora do âmbito da concorrência, não obstante a importância dos objetivos que visam prosseguir.

II. INTELIGÊNCIA ARTIFICIAL GENERATIVA

A IA Generativa é a inteligência artificial capaz de produzir conteúdos a partir de exemplos (cf. Figura 1). Engloba um conjunto diversificado de modelos estatísticos que são “treinados” para gerar dados semelhantes aos que constam de uma base de dados de treino (cf. Caixa 2). Os conteúdos gerados por IA Generativa são novos, não meras cópias dos exemplos na base de dados de treino.

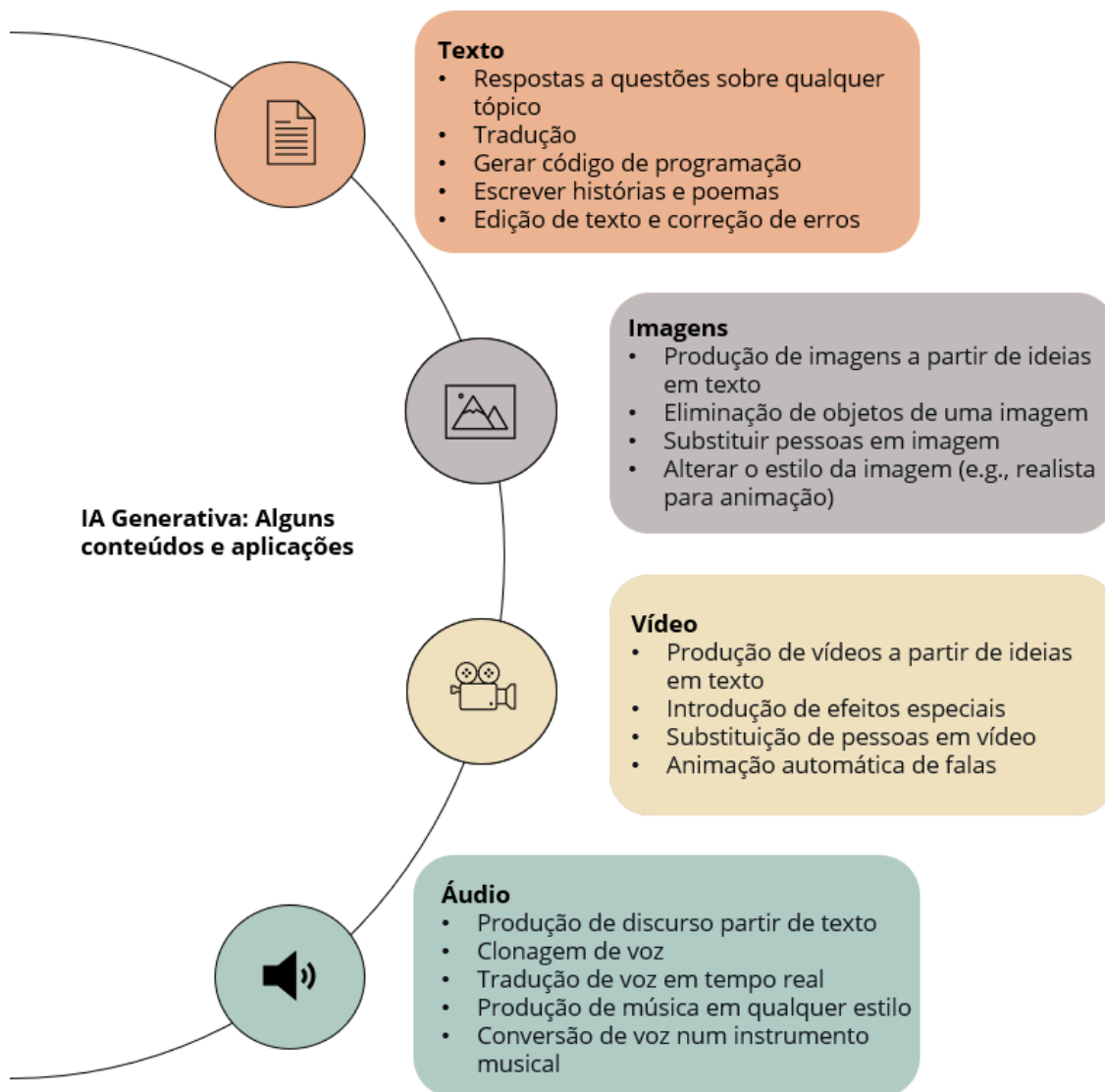
Figura 1 – Exemplos constantes de uma base de dados de treino utilizada para gerar conteúdos de texto



Fonte: Pedido à direita da imagem efetuado no serviço do ChatGPT (GPT-4), a 26.10.2023.

A IA Generativa pode gerar diversos tipos de conteúdos, incluindo texto, imagem, vídeo ou áudio (cf. Figura 2). Tem como único requisito os conteúdos poderem ser codificados em formato numérico. Cada tipo de conteúdo(s) requer o desenvolvimento de modelos estatísticos específicos, adaptados a esse(s) conteúdo(s).

Figura 2 – Exemplos de conteúdos produzidos por uma IA Generativa



Fonte: AdC.

Um mesmo modelo de IA Generativa pode combinar múltiplos tipos de conteúdo. Estes modelos, chamados de multimodais, estabelecem relações entre diferentes tipos de conteúdo, relacionando simultaneamente, por exemplo, texto, imagem e áudio. Assim, os pedidos dos utilizadores podem ser feitos em múltiplos formatos, e as respostas da IA Generativa podem ser igualmente produzidas em múltiplos formatos. Por exemplo, é possível fazer dobragem automática de filmes, o que combina clonagem de voz, tradução de voz e animação automática de fala.¹³ Os modelos multimodais diluem as diferenças entre os tipos de IA Generativa. Lidando com mais de uma modalidade ou tipo de dados, os

¹³ E.g., [Video Translate da HeyGen Labs](#).

modelos são treinados para compreender e gerar informações em vários formatos sem a necessidade de utilizar um conjunto de modelos especializados.¹⁴

Algumas aplicações de IA Generativa incluem:

- **A produção de texto a partir de um pedido em texto**, baseada em grandes modelos de linguagem (*Large Language Models*), como o GPT-3¹⁵ ou o GPT-4, incluído no ChatGPT da OpenAI.¹⁶ Refiram-se ainda o LaMDA, utilizado no Bard da Google,¹⁷ ou os LLaMA da Meta.¹⁸ Este tipo de modelos é capaz de gerar uma grande variedade de tipos de texto, incluindo texto expositivo, poesia ou código de programação.
- **A produção de imagens a partir de um pedido em texto ou imagens de exemplo** (cf. Figura 3), baseada em modelos de difusão (*Diffusion Models*¹⁹). Implementações disponíveis incluem o DALL-E 3 da OpenAI,²⁰ o Stable Diffusion da Stability AI²¹ e o Midjourney.²²
- **A produção de vídeos a partir de um pedido em texto, imagens ou vídeos de exemplo**, também baseada em modelos de difusão. O GEN-2 da Runaway é um exemplo de uma das implementações já disponíveis.²³
- **A produção de áudio a partir de um pedido em texto**. Um exemplo é o serviço de clonagem de voz fornecido pela Eleven Labs.²⁴ Estão também disponíveis serviços de IA Generativa que produzem música, como o MusicLM da Google.²⁵

¹⁴ E.g., Meta Blog (2023) "[ImageBind: Holistic AI learning across six modalities](#)", acessido a 18.10.2023; Google Blog (2023). "[What is Multimodal Search: 'LLMs with vision' change businesses](#)", acessido a 18.10.2023.

¹⁵ E.g., OpenAI (2020). Language Models are Few-Shot Learners, disponível [aqui](#).

¹⁶ OpenAI, "[GPT-4 is OpenAI's most advanced system, producing safer and more useful responses](#)", acessido a 18.10.2023. Cf. ainda OpenAI (2023). GPT-4 Technical Report, disponível [aqui](#).

¹⁷ Google Blog (2021) "[LaMDA: our breakthrough conversation technology](#)", acessido a 18.10.2023. Google Blog (2023) "[An important next step on our AI journey](#)", acessido a 18.10.2023.

¹⁸ Meta Blog (2023) "[Introducing LLaMA: A foundational, 65-billion-parameter large language model](#)", acessido a 18.10.2023. Meta Blog (2023) "[LLaMA 2: Open Foundation and Fine-Tuned Chat Models](#)", acessido a 18.10.2023.

¹⁹ E.g., Ho et al. (2020). Denoising Diffusion Probabilistic Models, disponível [aqui](#).

²⁰ OpenAI, "[Dall-E 3](#)", acessido a 18.10.2023.

²¹ Stability AI, "[Stable Diffusion XL](#)", acessido a 18.10.2023.

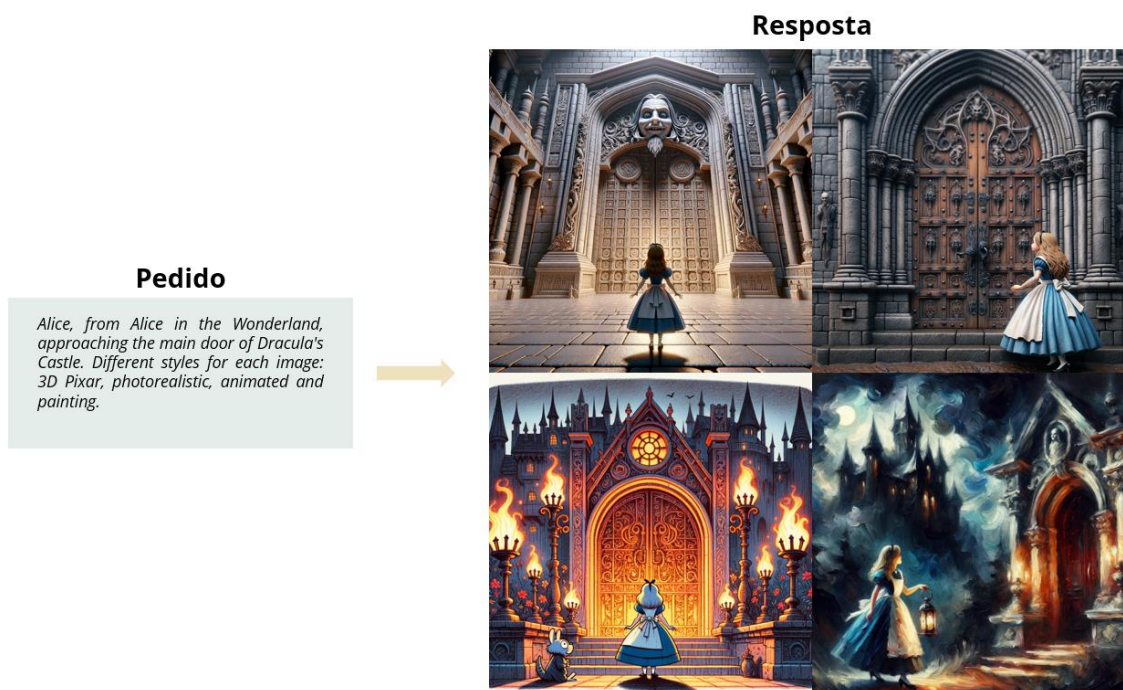
²² Midjourney, <https://www.midjourney.com/>, acessido a 18.10.2023.

²³ Runaway, "[Gen-2: The Next Step Forward for Generative AI](#)", acessido a 18.10.2023.

²⁴ Eleven Labs "[Generative Voice AI](#)", acessido a 18.10.2023.

²⁵ Google (2023). "[MusicLM: Generating music from text](#)".

Figura 3 – Imagem gerada por um serviço de IA Generativa a partir de um pedido de texto



Fonte: Pedido efetuado no plugin do Dall-E 3, no ChatGPT, a 26.10.2023.

As funcionalidades de uma IA Generativa podem ser estendidas através de *plugins*. Estes podem ser utilizados para combinar múltiplos serviços de IA Generativa, em que uma IA faz pedidos a outra. Os *plugins* também permitem conectar a IA Generativa a outros serviços menos “inteligentes”, como servidores de bases de dados ou interpretadores de programação.²⁶

Muitos destes serviços de IA Generativa encontram-se ainda em fase experimental. Os seus conteúdos apresentam por vezes pequenos defeitos que indiciam que foram gerados por IA Generativa. Entre estes refiram-se, no caso da geração de texto, as “alucinações”, em que a IA Generativa produz informação falsa, ainda que plausível à primeira vista. Os conteúdos gerados de imagens, vídeo e áudio, por seu turno, podem conter artefactos, como sejam detalhes incorretos,²⁷ movimentos pouco naturais de objetos,²⁸ ou a produção de sons poucos naturais.²⁹

²⁶ Os interpretadores de programação permitem executar código de programação gerado por uma IA Generativa, em resposta a pedidos do utilizador. E.g., Interpretador de Python no ChatGPT (GPT-4).

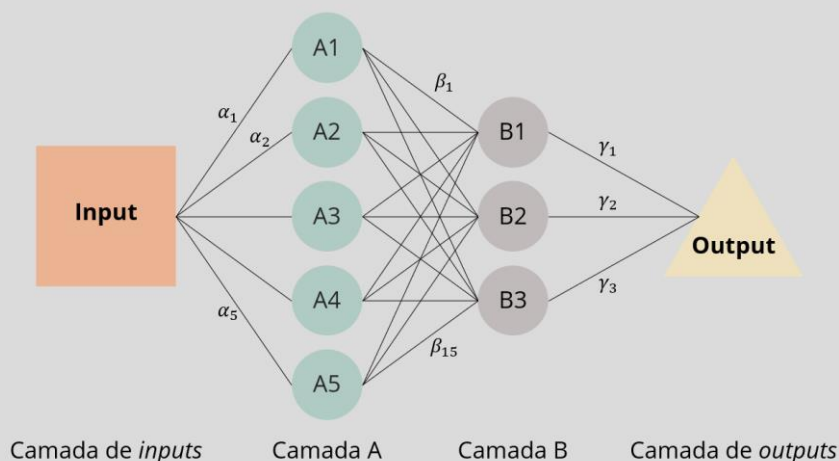
²⁷ E.g., na geração de uma imagem de uma pessoa, esta tem o número errado de dedos. Outro exemplo é a incapacidade de a IA Generativa de imagens produzir texto legível nas imagens.

²⁸ E.g., na transição de *frames* de um vídeo, haver mudanças repentinas de cor ou da forma dos objetos.

²⁹ E.g., Google (2016). WaveNet: A Generative Model for Raw Audio, disponível [aqui](#).

Não obstante, **tem-se verificado uma melhoria significativa da qualidade do conteúdo gerado**.³⁰ Embora a IA Generativa possa nunca vir a replicar a qualidade de conteúdo equiparável ao produzido por um ser humano especializado, permite uma aproximação em termos de qualidade, a um custo muito inferior e a uma velocidade de produção superior. Nessa medida, a IA Generativa poderá ter um impacto muito significativo, transversal a toda a economia, nomeadamente na produção de conteúdos de forma massificada.

Caixa 2 – Redes neuronais artificiais³¹



A IA Generativa tem por base arquiteturas de **redes neuronais artificiais**. Estes são modelos estatísticos que recebem um conjunto de *inputs*, aplicam várias operações matemáticas lineares e não lineares, e produzem *outputs*.

É comum conceber-se uma rede neuronal por via de um grafo (v. figura *supra*), que liga *inputs*, camadas (que representam operações lineares e não lineares) e *outputs*. Na figura *supra*, as camadas A e B aplicam operações sobre o *input*. A camada A tem 5 unidades, a camada B tem 3 unidades.

As redes neuronais artificiais são regidas por **parâmetros**, por uma **arquitetura** e por **hiperparâmetros**.

Os parâmetros são os valores que são ajustáveis durante o “treino” do modelo. Cada parâmetro é representado por uma linha na figura *supra* (e.g., α_1 , β_{15} ou γ_2). Os *outputs* do modelo são comparados com observações reais numa base de dados de treino (i.e., os exemplos). De seguida, os parâmetros são ajustados, num processo de otimização, conforme a diferença entre os *outputs* e os exemplos, tornando o modelo progressivamente

³⁰ O portal [PapersWithCode](#) compila testes e *benchmarks* de avaliação de diversos modelos de IA Generativa. É possível verificar que, ao longo do tempo, os resultados nestes testes e *benchmarks* têm melhorado (e.g., modelos de [produção de imagens](#), de [resposta a questões](#) ou de [produção de música](#)).

³¹ Existem muitos recursos disponíveis sobre este tópico. Três exemplos disponíveis gratuitamente *online*, à data (18.10.2023), são Goodfellow et al. (2016). Deep Learning, disponível [aqui](#); Zhang et al. (2021). Dive into Deep Learning, disponível [aqui](#); e Prince (2023). Understanding Deep Learning, disponível [aqui](#).

mais capaz de gerar dados semelhantes aos exemplos. Estes passos são o que se designa de “treino”.

Assim, durante o treino, o modelo sintetiza a informação contida nos dados e guarda-a nos seus parâmetros (*embeddings*). Os ficheiros que contêm esta informação constituem o modelo e são a coluna vertebral da IA Generativa.

A arquitetura e os hiperparâmetros dizem respeito a todas as configurações e valores do modelo definidos pelo programador, não sendo ajustáveis durante o treino. Genericamente, associa-se arquitetura ao tipo, estrutura e design do modelo (i.e., a sua topologia), e os hiperparâmetros a valores que regem o treino e ou desempenho do modelo.

O número e o tipo de camadas (e.g., linear ou não linear) de uma rede neuronal e as unidades em cada camada são exemplos de arquitetura/hiperparâmetros (v. Caixa 3). O número de camadas determina o quão “profundo” é o modelo. Os denominados modelos “*Deep Learning*” têm um número elevado de camadas (e.g., o GPT-3 tem 96) e, geralmente, permitem produzir modelos estatísticos mais complexos.

O treino de um modelo termina quando o seu programador assim o decidir, por exemplo, por estar satisfeito com o desempenho do modelo, ou por ter definido previamente a duração do treino. A duração do treino é um dos hiperparâmetros do modelo.

III. REQUISITOS PARA O DESENVOLVIMENTO E IMPLEMENTAÇÃO DE IA GENERATIVA

Existe uma diversidade de modelos de IA Generativa, que diferem conforme o tipo de conteúdo que produzem e a sua escala – avaliada de acordo com o número de parâmetros, o volume de dados ou o número de camadas da rede neuronal artificial. Um modelo é um conjunto de ficheiros que contém uma arquitetura, hiperparâmetros e parâmetros³² (cf. Caixa 2). Uma vez que os modelos atualmente disponíveis podem conter entre milhares a biliões³³ de parâmetros, estes ficheiros podem ser de grandes dimensões, de alguns GB até vários TB.

Podem distinguir-se genericamente os seguintes passos no desenvolvimento e implementação de IA Generativa (cf. Figura 4):

1. O **pré-treino**, em que se desenvolve um **modelo-base** (*foundation-model*).
2. O **transfer learning**, um conjunto de técnicas que realizam um treino adicional a partir de um modelo-base, capitalizando a informação nele sintetizada. Inclui, nomeadamente, o ***fine-tuning***, que alinha o modelo com os objetivos do

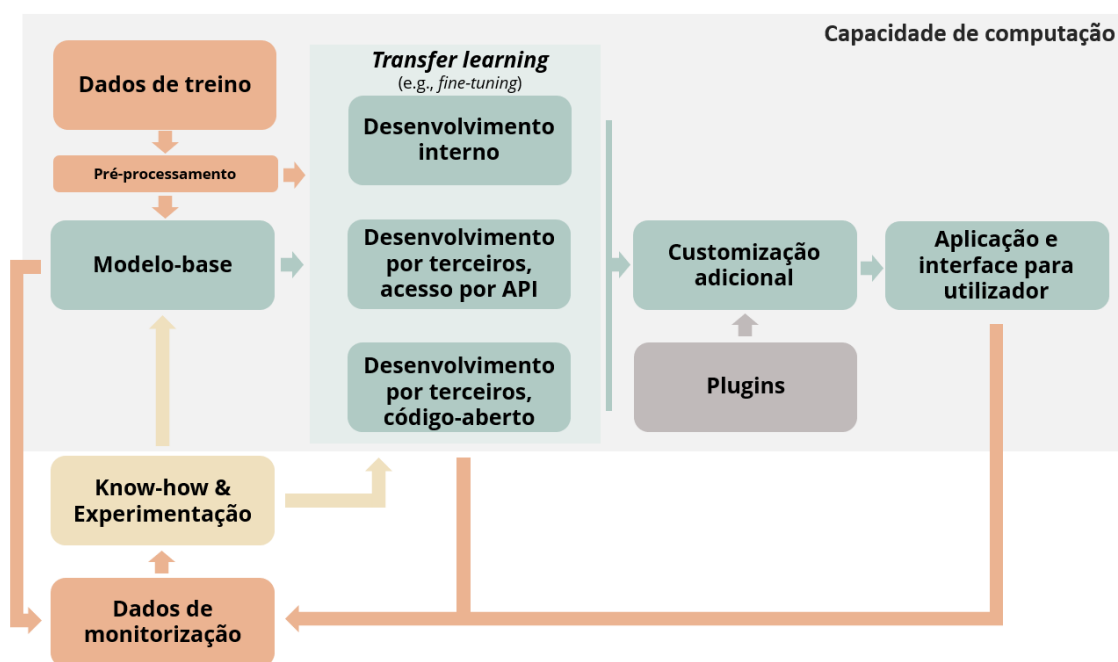
³² Os *model checkpoints* ou *model weights*.

³³ E.g. [este artigo](#) lista um conjunto de grandes modelos de linguagem com mais de 10 mil milhões de parâmetros treinados desde 2019. O maior modelo para o qual há informação disponível, [GLaM](#), tem 1,2 biliões de parâmetros. Cf. Zhao et al. (2023). A Survey of Large Language Models, p. 6. Adicionalmente, embora não seja informação confirmada pela OpenAI, alegadas fugas de informação sugerem que o GPT-4 terá quase 2 biliões de parâmetros. Cf. The decoder (2023) “[GPT-4 architecture, datasets, costs and more leaked](#)”, acedido a 18.10.2023. “Bilião” em escala longa, i.e., 10^{12} , equivalente ao “trillion” em inglês.

programador³⁴ e adapta o modelo a uma tarefa específica,³⁵ e o **knowledge distillation**, que comprime o modelo-base num modelo mais simples.³⁶

3. **Customização e parametrização adicionais**, que adaptam de forma mais completa o modelo a uma determinada tarefa, por exemplo, limitando o tipo de respostas que pode produzir. Também podem ser adicionados **plugins** à IA Generativa, que expandem as suas funcionalidades (cf. secção V.2).
4. **Implementação do serviço de IA Generativa direcionado a utilizadores** (e.g., aplicação ou interface).

Figura 4 – Cadeia de valor da IA Generativa



Fonte: AdC

³⁴ O alinhamento de IA refere-se ao problema de tentar construir IA que se comportem de acordo com os objetivos e as preferências de seres humanos. Cf., e.g., [esta entrevista a Paul Christiano, especialista em alinhamento de IA](#). Assegurar o alinhamento da IA é um dos passos do treino de modelo, sendo um tipo de *fine-tuning*. Cf., e.g., uma tentativa de resolver o problema de alinhamento, aplicada ao treino do GPT-3, disponível [aqui](#).

³⁵ No *fine-tuning*, partindo de um modelo pré-treinado, realiza-se um treino adicional, mantendo uma proporção dos parâmetros do modelo pré-treinado “congelados”. Não é necessário atualizar esses parâmetros, uma vez que já sintetizaram informação durante o treino do modelo pré-treinado. Os parâmetros “não congelados” vão sintetizar nova informação contida na base de dados de treino do modelo de *fine-tuning*, adaptando-o a uma tarefa específica.

³⁶ *Knowledge distillation* consiste em treinar um modelo mais simples a partir de outro mais complexo. Esta técnica é utilizada, em particular, para obter modelos menos exigentes do ponto de vista computacional e que, ainda assim, sintetizem muita da informação contida no modelo mais complexo.

Exemplificando, um banco pode pretender introduzir um serviço de *chat* inteligente para apoio ao cliente na sua página de Internet. Neste caso, o modelo-base seria o grande modelo de linguagem especializado em produzir texto. O modelo de *fine-tuning* seria o serviço de *chat* que responde a perguntas, adaptado a partir do modelo-base. A customização adicional limitaria o escopo de respostas possíveis pelo *chat* inteligente, por exemplo, não referindo nas respostas assuntos não relacionados com o banco ou mantendo sempre um tom de conversação apropriado com o cliente.³⁷ A implementação do serviço final consiste em criar a interface ou a aplicação, com a qual os clientes do banco vão interagir.

Estes passos são cumulativos, numa dependência vertical. Apenas é feito um treino de *fine-tuning* num modelo-base já sujeito a pré-treino. De igual modo, a implementação final do serviço de IA Generativa pressupõe que o modelo já se encontra adaptado a uma tarefa específica.

As fases de desenvolvimento e implementação de IA Generativa podem ser desempenhadas pelos mesmos operadores ou por operadores distintos. É possível, por exemplo, fazer *fine-tuning* a um modelo-base desenvolvido e treinado por terceiros (cf. secção V).

O desenvolvimento e a implementação de IA Generativa requerem acesso a dados, a capacidade de computação e a *know-how*. As próximas secções detalham cada um destes requisitos individualmente.

III.1.Dados

Podem distinguir-se dois tipos de dados relevantes para modelos de IA Generativa:

1. As **bases de dados de treino**, compostas pelos exemplos que treinam o modelo;
2. Os **dados de monitorização** do treino e do desempenho do modelo.

Bases de dados de treino

O treino de modelos de IA Generativa exige grandes volumes de dados, particularmente o treino de modelos-base. As técnicas de *transfer learning*, como o *fine-*

³⁷ E.g., o *plugin* da Expedia – agência de viagens – no ChatGPT submete ao ChatGPT as seguintes instruções: "In ALL responses, Assistant MUST always start with explaining assumed or default parameters. In addition, Assistant MUST always inform user it is possible to adjust these parameters for more accurate recommendations. Assistant explains its logic for making the recommendation. Assistant presents ALL the information within the API response, especially the complete Expedia URLs to book in markdown format. For each recommended item, Assistant always presents the general descriptions first in logical and readable sentences, then lists bullets for the other metadata information. Assistant encourages user to be more interactive at the end of the recommendation by asking for user preference and recommending other travel services. Here are two examples, "What do you think about these? The more you tell me about what you're looking for, the more I can help!", "I'd like to find a trip that's just right for you. If you'd like to see something different, tell me more about it, and I can show you more choices.". Assistant must NEVER add extra information to the API response. Assistant must NEVER mention companies other than Expedia or its sub-brands when relaying the information from Expedia plugin." Cf. [ChatGPT Plus](#) e [artigo informativo sobre o plugin Expedia no ChatGPT](#), acessado a 27.10.2023.

tuning, apoiando-se num modelo-base pré-treinado, utilizando um menor volume de dados. Ainda assim, o tipo de dados usado em modelos-base e modelos de *transfer learning* é diferente. Modelos de *fine-tuning*, por exemplo, tendem a exigir bases de dados mais pequenas e curadas, muitas vezes criadas manualmente por seres humanos.

As bases de dados utilizadas no treino de modelos de IA Generativa têm origens diversas, e podem exigir custos de aquisição. Muitos dos modelos atualmente disponíveis são treinados com recurso a bases de dados publicamente disponíveis e estruturadas. Ainda assim, é comum conjugarem-nas com dados não estruturados (e.g., obtidos por *web scraping*), ou com dados não acessíveis ao público, sejam privados ou sujeitos a licenciamento. Estes dados privados podem exigir um custo adicional de aquisição.

O GPT-3 da OpenAI, por exemplo, foi treinado com recurso a uma combinação de arquivos públicos da Internet e arquivos de livros,³⁸ num total de 300 mil milhões de *tokens*.³⁹ No caso do GPT-4, a OpenAI não revelou as bases de dados utilizadas.⁴⁰ Por sua vez, o primeiro modelo LLaMA da Meta recorre inteiramente a bases de dados públicas,⁴¹ num total de 1,4 biliões de *tokens*. Já um dos modelos Stable Diffusion da Stability AI foi treinado a partir da base de dados de pares texto-imagem retirada de um arquivo da Internet, com 400 milhões de pares texto-imagem.⁴²

A diversidade e qualidade dos dados de treino utilizados são importantes para treinar um modelo com bom desempenho. A informação pública disponível sobre os atuais modelos não permite concluir sobre a importância relativa das bases de dados, quando várias são utilizadas conjuntamente. No entanto, os modelos tendem a recorrer a múltiplas bases de dados, não obstante algumas terem uma dimensão reduzida, de modo a tirar partido da capacidade de generalização dos modelos.⁴³ A necessidade de bases de dados de qualidade pode implicar custos acrescidos de obtenção de dados.

Poderão existir efeitos de rede na obtenção de dados, caso os serviços de IA Generativa incluam funcionalidades de *upload* de ficheiros por utilizadores. Caso o fornecedor de IA Generativa tenha a capacidade e a permissão, os dados submetidos pelos

³⁸ O GPT-3 foi treinado com recurso aos arquivos de Internet [Common Crawl](#), o [WebText2](#) e a [Wikipédia inglesa](#). A OpenAI não especifica a origem das bases de dados de livros que utiliza, não sendo claro qual a natureza da base de dados. A maioria dos dados de treino tem origem no Common Crawl. Cf. OpenAI (2020). Language Models are Few-Short Learners (pp. 8-9).

³⁹ O *token* é a unidade de dados em grandes modelos de linguagem, i.e., cada *token* é uma observação. Genericamente, cada *token* representa uma palavra ou uma parte de uma palavra.

⁴⁰ OpenAI (2023). GPT-4 Technical Report (p. 2).

⁴¹ E.g., arquivos de Internet: Common Crawl, C4, Github, Wikipedia, Stack Exchange. Arquivos de livros: Gutenberg e ThePile. Arquivo de ficheiros: LaTeX do repositório de *papers* ArXiv. Cf. Meta AI (2023). LLaMA: Open and Efficient Foundation Language Models (p. 2).

⁴² [Base de dados LAION](#), construída a partir do arquivo Common Crawl, extraíndo as imagens e o texto associado a cada imagem. Cf. Rombach et al. (2021). High-Resolution Image Synthesis with Latent Diffusion Models, disponível [aqui](#).

⁴³ E.g., no caso de modelos de texto, Zhao et al. (2023). A Survey of Large Language Models, pp. 9-13.

utilizadores podem ser convertidos em dados de treino. Neste cenário, quanto maior a base de utilizadores, mais facilitado o acesso a novos dados de treino.

Os dados, antes de serem utilizados em modelos de IA Generativa, são tipicamente sujeitos a um pré-processamento. Conforme o tipo de conteúdo produzido pela IA Generativa, os dados têm características diversas, que determinam como podem ser pré-processados. Independentemente de os exemplos de treino serem texto, imagem, vídeo, áudio, ou outro, a informação é convertida num formato numérico estandardizado para poder ser processada pelo modelo.⁴⁴

O pré-processamento de dados utilizado pode ter impacto no desempenho do modelo final.⁴⁵ Muitas bases de dados misturam informação de qualidade variável.⁴⁶ Dado o impacto da qualidade dos dados no desempenho do modelo, um dos principais objetivos do pré-processamento é filtrar os dados de boa qualidade. Os mesmos dados de treino podem ser pré-processados de formas diferentes, pelo que o pré-processamento aplicado é uma escolha dos programadores do modelo e a sua otimização requer experimentação (cf. secção III.3 e Caixa 3).

Dados de monitorização

O desenvolvimento de modelos de IA Generativa é um processo que requer muita experimentação por parte dos programadores, de modo a otimizar o processo de treino e o desempenho dos modelos. Em particular, a experimentação serve para escolher de forma ótima a arquitetura e os hiperparâmetros do modelo (cf. secção III.3).

Assim, os dados de monitorização do treino e do desempenho podem ser um input essencial no desenvolvimento e implementação de modelos de IA Generativa. Não obstante vários dos modelos atualmente disponíveis recorrerem a bases de dados de treino públicas, o mesmo não sucede para os dados de monitorização. Um resumo desta informação, ainda assim, é geralmente reportado nas publicações e artigos que apresentam detalhes técnicos dos modelos.⁴⁷

A natureza privada dos dados de monitorização tornar-se-á mais relevante num contexto de implementação de serviços de IA Generativa em massa, para o utilizador final. Nesse cenário, o fornecedor de serviços de IA Generativa pode recolher grandes volumes de dados sobre utilizadores finais, ganhando acesso a uma base de dados privada

⁴⁴ Se o conteúdo for imagens, por exemplo, estas são tipicamente convertidas para ficar com a mesma dimensão ou com o mesmo esquema de cores. Por outro lado, o texto será convertido em representações numéricas que permitem, uma vez treinado o modelo, relacionar palavras com significados ou contexto semelhantes (e.g., a relação de “homem” para “mulher” é a mesma que a de “rei” para “rainha”).

⁴⁵ No caso de modelos de texto, Zhao et al. (2023). A Survey of Large Language Models, pp. 5,16-17.

⁴⁶ Um exemplo é a base de dados [Common Crawl](#), utilizada no treino do GPT-3 e do LLaMA.

⁴⁷ Veja-se, e.g., o caso do GPT-3, para o qual a OpenAI indica alguma informação sobre a arquitetura e hiperparâmetros dos modelos que treinou (pp. 8, 43), assim como as necessidades de computação do treino (p. 9) e as curvas de treino (pp. 10-11). OpenAI (2020). Language Models are Few-Short Learners.

que permite monitorizar o desempenho do modelo de IA Generativa a partir do comportamento e *feedback* dos utilizadores.

O desenvolvimento e a implementação de IA Generativa pode beneficiar de efeitos de rede, na medida em que o desempenho do modelo e assim, o respetivo valor, aumente com o número de utilizadores. Conforme se detalhará *infra*, estes efeitos de rede podem conferir uma vantagem competitiva ao maior fornecedor do mercado (cf. secção IV.2).

É possível conceber várias formas como um fornecedor de IA Generativa pode observar o comportamento ou obter *feedback* dos utilizadores e, com isso, monitorizar o desempenho do modelo e aumentar a qualidade e o grau de personalização da IA Generativa.⁴⁸ Por exemplo:

- Pode-se monitorizar se os utilizadores, no seguimento de um pedido inicial, fazem pedidos subsequentes ligeiramente diferentes para clarificar o pedido inicial. Este comportamento indicia que o utilizador não ficou inteiramente satisfeito com a resposta da IA Generativa ao pedido inicial e que as respostas subsequentes são preferíveis.
- Pode-se monitorizar se o utilizador copia algum tipo de conteúdo gerado pela IA Generativa, o que seria tomado como indício que a IA Generativa produziu conteúdo do seu agrado.

A monitorização dos utilizadores pode ser combinada com metodologias de testes A/B.⁴⁹ Estes testes podem apresentar a uma minoria de utilizadores variantes alternativas do modelo de IA Generativa, permitindo testar diferentes arquiteturas e hiperparâmetros, avaliando-as de acordo com métricas relevantes (cf. secção III.3), e com isso realizar uma experimentação mais intensiva de modelos. É possível correr vários testes A/B ou testes com várias variantes em simultâneo.

III.2. Capacidade de computação

O treino de modelos de IA Generativa, especialmente modelos-base, requer uma capacidade de computação significativa, exigindo *hardware* especializado, nomeadamente supercomputadores e sistemas de computação distribuída, com um grande número de CPU e GPU.⁵⁰ Os modelos GPT da OpenAI, por exemplo, são treinados na *cloud*

⁴⁸ E.g., A OpenAI refere, [neste FAQ](#), que o comportamento dos utilizadores do ChatGPT é monitorizado para melhorar o desempenho do modelo, na secção “*How does OpenAI use my personal data?*”. Adicionalmente, [neste artigo](#), a OpenAI descreve como monitoriza o funcionamento do ChatGPT para o melhorar, no que respeita a segurança dos conteúdos produzidos e a prevenção de abusos por parte de utilizadores.

⁴⁹ Os testes A/B apresentam aleatoriamente aos utilizadores variantes de alguma funcionalidade de um serviço digital, e.g., modelos com uma arquitetura ou hiperparâmetros diferentes, no caso da IA Generativa. As variantes são depois comparadas de acordo com uma métrica relevante, previamente definida.

⁵⁰ CPU: Central Processing Unit (vulgo “processador”). GPU: Graphical Processing Unit (vulgo “placa gráfica”).

da Microsoft, que em 2020 desenvolveu um supercomputador exclusivamente para a OpenAI com 285 000 núcleos de CPU e 10 000 GPU.⁵¹

Este treino pode durar até vários meses, ainda que seja possível reduzir o tempo de treino recorrendo a computação paralela (e.g., utilizando um grande número de GPU em simultâneo). Veja-se, por exemplo, o LLaMA da Meta, que durou 21 dias⁵², ou o GPT-3 que se estima que terá durado pouco mais de um mês.^{53,54}

A implementação de serviços de IA Generativa⁵⁵ requer também uma capacidade de computação significativa, conforme a maior dimensão dos modelos, exigindo *hardware* dedicado.

Não obstante, **é tecnicamente possível correr modelos de IA Generativa em *hardware* menos sofisticado**, como computadores pessoais comuns,⁵⁶ nomeadamente modelos de *transfer learning*. De facto, existem várias iniciativas, que procuram treinar modelos de IA Generativa alternativos, a custos inferiores e recorrendo a capacidades de computação mais modestas, nomeadamente em formato de código aberto.⁵⁷ Em contrapartida, estes modelos sacrificam a dimensão do modelo e ou pode ser necessário esperar algum tempo para que a IA Generativa responda a pedidos.

Os custos de treino e de implementação de um modelo dependem do tipo de *hardware* utilizado (ou, em alternativa, do custo dos serviços de computação na *cloud*), da sua arquitetura e hiperparâmetros (e.g., número de camadas e de parâmetros do modelo, duração do treino), do volume de dados e da extensão da experimentação pelo programador.

É difícil avaliar a dimensão dos custos associados ao desenvolvimento e à implementação de modelos de IA Generativa. No entanto, é possível obter estimativas considerando as necessidades de computação reportadas pelas publicações que

⁵¹ OpenAI (2020). Language Models are Few-Short Learners (p. 9); OpenAI (2023). GPT-4 Technical Report (p. 17); Microsoft News (2020) "[Microsoft announces new supercomputer, lays out vision for future AI work](#)", acedido a 20.10.2023; Microsoft Blog (2023) "[Microsoft and OpenAI extend partnership](#)", acedido a 20.10.2023; OpenAI Blog (2023) "[OpenAI and Microsoft extend partnership](#)", acedido a 20.10.2023.

⁵² Meta AI (2023). LLaMA: Open and Efficient Foundation Language Models. (p. 4).

⁵³ Narayanan et al. (2021) Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM (p. 8).

⁵⁴ Ver ainda outros exemplos de grandes modelos de linguagem, em Zhao et al. (2023). A Survey of Large Language Models, p. 6.

⁵⁵ Por "implementação de serviços de IA Generativa" entende-se a utilização dos modelos para fazer inferência, i.e., gerar *outputs* a partir de *inputs*, que acontece sempre que um utilizador faz um pedido (o *input*) e a IA Generativa produz uma resposta (o *output*).

⁵⁶ Em março de 2023, foi publicado um modelo treinado por *fine-tuning* a partir do LLaMA da Meta, denominado Alpaca – descrição disponível [aqui](#). A partir deste modelo foram criadas várias implementações que permitem correr a versão mais simples (i.e., usando o modelo-base mais pequeno do LLaMA) em qualquer computador – e.g., [Alpaca Electron](#).

⁵⁷ [Esta página](#) elenca alguns dos grandes modelos de linguagem disponibilizados de forma aberta.

introduzem alguns dos principais modelos, assim como as necessidades de computação para responder a cada pedido de um utilizador.⁵⁸

Ressalva-se que os custos de acesso a capacidade de computação podem alterar-se no tempo, sendo menores com inovação tecnológica ou, possivelmente, maiores caso haja restrições de oferta ou aumentos de procura (e.g., grande procura por GPU especializada em IA Generativa).⁵⁹ Os custos também podem reduzir-se com a maior eficiência dos modelos, que mantêm o desempenho apesar de terem sido treinados com menos dados ou durante menos tempo. Adicionalmente, apesar de se tomar como referência os custos de acesso a capacidade de computação na *cloud* para estimar custos, certos operadores podem beneficiar de custos mais baixos.

É possível que a IA Generativa não esteja apenas ao alcance de um reduzido número de empresas, com avultadas capacidades de computação e financeiras. Todavia, o nível de experimentação necessário para o treino de novos modelos pode aumentar significativamente os custos de desenvolvimento.

Em primeiro lugar, um modelo, quando divulgado ao público, inclui várias configurações diferentes,⁶⁰ que se distinguem, nomeadamente pelo número de parâmetros e pelo volume da base de dados de treino. Por exemplo, o GPT-3 foi disponibilizado em oito variantes distintas⁶¹ e o LLaMA em quatro. Cada um destas variantes do modelo foi sujeito a um treino independente, multiplicando os custos de treino.

Em segundo lugar, os modelos divulgados publicamente não são necessariamente todos os modelos treinados pelo fornecedor de IA Generativa. A experimentação visa escolher uma arquitetura e hiperparâmetros ótimos do ponto de vista da eficiência de treino e do desempenho do modelo. Os modelos divulgados resultam, assim, já de um processo de seleção entre vários modelos (cf. secção III.3).

Em terceiro lugar, estar na fronteira tecnológica pode implicar uma maior intensidade de experimentação, assim como maior capacidade (e possivelmente custos) de computação.

⁵⁸ Têm sido reportadas muitas estimativas sobre os custos de desenvolver e implementar IA Generativa. Mais que valores exatos, a partir estimativas de fontes diferentes (incluindo, por vezes, os próprios fornecedores), é possível ter uma ideia da ordem de magnitude dos custos. Sobre os custos de desenvolvimento (i.e., de treino), tem sido, por exemplo, estimado que [o treino do GPT-3 terá custado 4 milhões de dólares \(ou quase 5 milhões, noutra estimativa\)](#) e que o treino do [GPT-4 terá custado mais de 100 milhões de dólares](#). Por seu turno, [o Stable Diffusion terá custado 600 mil dólares](#). Relativamente aos custos de implementação dos modelos de IA Generativa, tem sido estimado que [o ChatGPT estará a custar cerca de 700 mil dólares por dia](#), ou [alguns centimos por chat](#). Exemplos de outros artigos que discutem os possíveis custos dos modelos de IA Generativa e a sua evolução, [aqui](#), [aqui](#) e [aqui](#).

⁵⁹ Ver nota de rodapé anterior e secção IV.3.

⁶⁰ Tipicamente, quando se refere ao modelo sem mais nenhum qualificativo, refere-se à variante do mais complexa.

⁶¹ OpenAI (2020). Language Models are Few-Short Learners (p. 8).

Por estes motivos, as estimativas do custo de treino de um único modelo podem ser enganadoras quanto à capacidade de outros operadores no mercado treinarem modelos semelhantes. Neste contexto, veja-se que o CEO da OpenAI afirmou publicamente que treinar o GPT-4 custou mais de 100 milhões de dólares.⁶² Não é, porém, claro se tal valor se refere ao treino do modelo final do GPT-4 ou a todo o processo de desenvolvimento.

III.3. Know-how e experimentação

O desenvolvimento de modelos de IA Generativa requer muita experimentação, com vista a otimizar a arquitetura e os hiperparâmetros do modelo, para que, os custos do treino sejam baixos, a velocidade do treino seja rápida e o modelo resultante, uma vez finalizado o treino, tenha um desempenho elevado.

O desenvolvimento destes modelos está sujeito a um grande número de graus de liberdade, exigindo, por parte do programador, um *know-how* significativo. O programador tem de tomar muitas pequenas decisões que podem ser determinantes, cumulativamente e, por vezes, individualmente, para a competitividade do modelo final de IA Generativa (cf. Caixa 3).

⁶² Wired (2023) "[OpenAI's CEO Says the Age of Giant AI Models Is Already Over](#)", acedido a 22.10.2023.

Caixa 3 – Arquitetura e hiperparâmetros⁶³

Lista ilustrativa de algumas das escolhas feitas pelo programador quanto à arquitetura e aos hiperparâmetros de um modelo, numa rede neuronal artificial:

- Base de dados de treino: O tipo de amostragem e a dimensão da base de dados.
- Pré-processamento: O tipo de normalização aplicada, o número de *embeddings*, as técnicas de *data augmentation* (e.g., em imagens, a rotação, as alterações de dimensões, a simetria, o *zoom*, a deslocação, o *crop*, etc.). A eliminação de observações de má qualidade ou de dados duplicados, a definição de “exemplos positivos” e de “exemplos negativos”.
- Camadas intermédias: O número e tipo de camadas intermédias, o número de unidades em cada camada, e a sua ordem e combinação. Exemplos de tipos de camadas intermédias: *linear*, *ativação*, *convolution*, *pooling*, *transformer*, *flatten*, *dropout*, etc. Cada camada implica também escolhas, e.g. na camada de ativação é necessário escolher a função de ativação (e.g., ReLU ou sigmoide); na camada de *convolution* é necessário escolher o *kernel*, o *stride* e o *padding*. As escolhas quanto a camadas intermédias determinam o número de parâmetros do modelo.
- Treino: O número de *epochs* do treino, i.e., quando parar o treino. O número de *batches* no treino e o método de amostragem dos *batches*. O algoritmo de otimização utilizado no método dos gradientes. A taxa de aprendizagem (*learning rate*) e a forma como evolui no decurso do treino (e.g., *learning rate scheduler*, *learning rate warmup*). O tipo de função de custo utilizado no treino. O *weight decay*, o *label smoothing* e as técnicas de regularização aplicadas.
- Transfer learning: A escolha do modelo-base, a escolha das camadas cujos parâmetros são “congelados” (no caso de *fine-tuning*) ou a escolha da arquitetura e hiperparâmetros do modelo de *knowledge distillation*.

A escolha de arquitetura e de hiperparâmetros não é geralmente um processo simples, e dependerá, em larga medida, da experimentação, da experiência e *know-how* dos programadores, e das melhores práticas do setor. Com efeito, não parecem existir critérios teóricos bem definidos para a escolha ótima de arquitetura ou de hiperparâmetros, sendo reconhecido que esta escolha é uma “arte”.⁶⁴

As ferramentas de desenvolvimento de IA Generativa incluem, de modo geral, diversas funcionalidades para fazer e monitorizar experimentação (*experiment tracking*).⁶⁵ Estes

⁶³ Informação compilada a partir das referências apresentadas *supra*. Cf. nota de rodapé 31.

⁶⁴ O treino de modelos de IA Generativa é regularmente descrito como sendo uma “arte”, dado não haver métodos teóricos consolidados sobre como escolher a arquitetura e os hiperparâmetros de um modelo. E.g., Waring et al. (2020). *Automated machine learning: Review of the state-of-the-art and opportunities for healthcare*; ou Wolfram (2023) “[What Is ChatGPT Doing ... and Why Does It Work?](#)”, acessado a 20.10.2023. A escolha de hiperparâmetros é geralmente referida como “*hyperparameter optimisation*” ou “*hyperparameter tuning*”. Ver mais, e.g., em Zhang et al. (2021) *Dive into deep learning* (Cap. 19 [Hyperparameter Optimization](#)).

⁶⁵ E.g., Tensorboard, disponível [aqui](#).

podem vir a ser complementados com experimentação direta junto dos utilizadores finais, recorrendo a testes A/B.

III.4. Outros aspetos necessários para o desenvolvimento e implementação de IA Generativa

A IA Generativa levanta um conjunto alargado de questões, nomeadamente relacionadas com privacidade, segurança e propriedade intelectual, passíveis de ter impacto no seu desenvolvimento e implementação no mercado. Estas incluem questões de direitos de autor relacionadas com o recurso a obras de autor para efeitos dos dados de treino⁶⁶, bem como questões de privacidade, incluindo a forma como os dados são divulgados e a capacidade de proporcionar direitos aos titulares dos dados. O recurso a grandes volumes de dados e o respetivo treino do modelo poderá também suscitar riscos de manipulação dos dados de treino com vista a influenciar as respostas do modelo de forma deliberada para produzir desinformação. Outras questões relacionam-se com a fiabilidade das respostas e a segurança do sistema de IA Generativa.⁶⁷

Estas questões têm suscitado discussões ao nível dos decisores públicos e organizações internacionais têm espoletado propostas legislativas e de instrumentos não vinculativos. A importância de assegurar sistemas de IA seguros já foi reconhecida na Recomendação do Conselho da OCDE em matéria de IA, de 2019.⁶⁸ Em particular, a OCDE recomenda que os governos devem rever e adaptar, se necessário e aplicável à IA, o respetivo enquadramento regulatório e de política pública com vista a promover a inovação e a concorrência no domínio de sistemas de IA confiáveis (cf. princípio 2.3 b). A nível da UE, importa destacar a Proposta de Regulamento IA que já estabelece, inclusive, obrigações específicas para fornecedores de modelos-base e fornecedores a jusante (cf. Caixa 4). A UE optou por uma abordagem baseada no risco para a governação da IA. A 30 de outubro de 2023, o Presidente Biden emitiu uma *Executive Order* sobre a IA com vista a promover a equidade e os direitos civis, defender os consumidores e os trabalhadores, promover a inovação e a concorrência, e assegurar a segurança dos sistemas de IA.⁶⁹

⁶⁶ Com efeito, já foram instauradas ações judiciais, nos Estados Unidos, contra o uso de obras protegidas por direitos de autor para treinar IA Generativa sem permissão prévia. E.g., [Chabon et al v. OpenAI, Inc. et al](#); [Tremblay et al v. OpenAI, Inc. et al](#).

⁶⁷ Para mais informação sobre estes riscos, cf. e.g., Bommasani et al. (2021). On the opportunities and risks of foundation models.

⁶⁸ Recomendação do Conselho da OCDE sobre Inteligência Artificial, adotada em 22 de maio de 2019 [OECD/LEGAL/0449], disponível [aqui](#).

⁶⁹ Disponível em [Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence | The White House](#).

Caixa 4 – Proposta de Regulamento IA da EU

Em abril de 2021, a CE apresentou uma proposta de Regulamento Inteligência Artificial, que estabelece regras harmonizadas em matéria de IA, tendo a mesma sido sujeita a alterações em junho de 2023.⁷⁰ A proposta visa: (i) garantir que os sistemas de IA são seguros e respeitam os direitos fundamentais; (ii) garantir a segurança jurídica; (iii) reforçar a governação e a aplicação efetiva da legislação existente; e (iv) facilitar o desenvolvimento de um mercado único para aplicações de IA seguras.

A proposta estabelece uma definição tecnologicamente neutra de sistemas de IA e classifica-os conforme o risco associado às diferentes utilizações da IA. A proposta prevê que: (i) serão proibidas práticas específicas de IA em contradição com os valores e direitos fundamentais da UE; (ii) os sistemas de IA de alto risco devem cumprir um conjunto de requisitos e os seus fornecedores e utilizadores devem estar sujeitos a um conjunto de obrigações; e (iii) determinados sistemas de IA que não sejam de alto risco devem cumprir obrigações de transparência.

A proposta inclui um conjunto de obrigações específicas que os fornecedores de modelos-base devem cumprir. Por exemplo, devem (i) tratar e incorporar apenas dados sujeitos a medidas adequadas de governação de dados; (ii) assegurar níveis adequados de desempenho; (iii) elaborar documentação técnica e instruções de utilização, a fim de permitir que os fornecedores a jusante cumpram as suas obrigações previstas na proposta de Regulamento; (iv) estabelecer um sistema de gestão de qualidade; e (v) registar o modelo-base numa base de dados da UE (cf. n.º 2 do art.º 28-B da proposta de alteração).

São ainda estabelecidas obrigações adicionais aos fornecedores de modelos-base e aos fornecedores a jusante. Estes devem (i) cumprir com obrigações de transparência; (ii) assegurar salvaguardas adequadas contra a geração de conteúdos em violação da legislação da UE; e (iii) disponibilizar ao público um resumo pormenorizado da utilização dos dados de treino protegidos pela legislação em matéria de direitos de autor (cf. n.º 4, art.º 28-B da proposta de alteração).

A proposta exclui do seu âmbito os componentes de IA prestados ao abrigo de licenças livres e em código aberto, à exceção dos modelos-base e outros sistemas de IA de risco elevado (cf. n.º 5-E, art.º 2.º da proposta de alteração).

A proposta prevê igualmente medidas de apoio à inovação, incluindo a promoção de *sandboxes* regulatórias para reduzir os encargos regulamentares e apoiar as PME e *start-ups* (cf. n.º 1 (e-A), art.º 1.º da proposta de alteração).

Prevê-se que o Regulamento IA seja complementado, a nível mundial, pelos princípios orientadores e pelo código de conduta voluntário do G7 para os prestadores de sistemas de IA. Em outubro de 2023, a CE realizou um inquérito a *stakeholders* sobre um projeto de princípios orientadores.⁷¹ O projeto prevê 11 princípios, incluindo, entre outros, a identificação de vulnerabilidades, a partilha responsável de informações e a comunicação de incidentes, a aplicação de controlos de segurança sólidos, o desenvolvimento de conteúdos fiáveis, a adoção de normas técnicas internacionais, se adequado, e a aplicação de controlos e auditorias adequados aos dados de entrada.

IV. CONCORRÊNCIA ENTRE FORNECEDORES DE MODELOS-BASE

IV.1. Concorrência em qualidade

Ainda não são claras quais serão as dimensões de concorrência no mercado de modelos-base, por o mercado ainda estar numa fase embrionária. Contudo, parece já possível notar que a qualidade será uma das dimensões-chave de concorrência entre modelos-base. A qualidade do produto deve ser avaliada à luz da capacidade de um produto para satisfazer as necessidades de um consumidor. Nessa ótica, ainda não haverá elementos suficientes para distinguir quais os fatores que determinam a qualidade de um modelo-base.

Não obstante, **o setor já concebeu várias metodologias para avaliar os modelos-base**, aplicadas na fase de avaliação do modelo, que podem servir de *proxy* para as diferenças de qualidade entre modelos. Estas metodologias podem ser mais ou menos automáticas, ou necessitar de mais ou menos intervenção humana, havendo, entre os dois formatos um *trade-off* entre o custo e rapidez de avaliação e a sua qualidade.

Entre as metodologias utilizadas, as mais comuns são um conjunto de métricas, testes e benchmarks padronizados no setor, e que permitem comparar diretamente o desempenho de modelos. Assim, na produção de texto, por exemplo, procura-se avaliar se a IA Generativa produz informação verdadeira,⁷² responde corretamente a questões sobre tópicos diversos,⁷³ é capaz de resolver problemas matemáticos⁷⁴ ou questões de senso comum⁷⁵ ou é capaz de gerar código.⁷⁶ Por seu turno, na produção de imagens, avalia-se a

⁷⁰ À data da redação, o processo ainda estava em curso. Em 14.06.2023, o Parlamento Europeu aprovou alterações sobre a proposta de regulamento ([Regulamento Inteligência Artificial](#)) que altera determinados atos legislativos da união ([COM\(2021\)0206](#) – C9-0146/2021 – [2021/0106\(COD\)](#)).

⁷¹ Cf., “International Draft Guiding Principles for Organizations Developing Advanced AI systems”, disponível [aqui](#).

⁷² E.g., o teste TruthfulQA. Este teste é composto por questões que muitas pessoas respondem incorretamente, seja devido a crenças falsas ou a algum erro de raciocínio. Lin et al. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. Disponível [aqui](#) e [aqui](#).

⁷³ E.g., o relatório técnico sobre o GPT-4 avalia o desempenho do modelo com vários testes académicos e de acesso a profissões, abrangendo áreas como direito, medicina, biologia, economia, história, matemática ou línguas – OpenAI (2023). GPT-4 Technical Report, pp. 4-6. Existem ainda *benchmarks* estandardizados que avaliam os modelos neste tipo de questões, como o MMLU – Hendrycks et al. (2021). Measuring Massive Multitask Language Understanding, disponível [aqui](#).

⁷⁴ E.g., o GSM8K é um *benchmark* que contém problemas matemáticos ao nível de escola primária. Cf. OpenAI (2021). Training Verifiers to Solve Math Problems, disponível [aqui](#). Adicionalmente, o MATH é um *benchmark* que contém problemas matemáticos ao nível do ensino secundário. Cf. Hendrycks et al. (2021). Measuring Mathematical Problem Solving with the MATH Dataset, disponível [aqui](#).

⁷⁵ E.g., o HellaSwag é um *benchmark* que avalia se o modelo é capaz de completar frases sobre situações banais de um modo que faça sentido. Cf. Zellers et al. (2019). HellaSwag: Can a Machine Really Finish Your Sentence?, disponível [aqui](#).

⁷⁶ E.g. o HumanEval avalia se o modelo é capaz de produzir código em Python. OpenAI (2021) Evaluating Large Language Models Trained on Code. Disponível [aqui](#) e [aqui](#).

proximidade da imagem produzida ao pedido de texto submetido ao modelo,⁷⁷ comparam-se imagens reais com imagens geradas pelo modelo de IA Generativa⁷⁸ ou estas são avaliadas diretamente por seres humanos.

A intervenção humana na avaliação de modelos também se pode estender aos dados de monitorização, como a observação do comportamento dos utilizadores dos modelos ou o *feedback* que fornecem. Esta informação pode ser utilizada para melhorar futuras iterações dos modelos.

Outras possíveis dimensões da qualidade do modelo resultam de escolhas diretas dos programadores. Por exemplo, no contexto de produção de texto, a memória (*contexto length*) do modelo é definida pelo programador e determina a quantidade de informação que o modelo é capaz de processar em simultâneo quando gera uma resposta. No âmbito de um *chatbot*, em particular, maior memória significa que a IA Generativa tem em conta um maior número de mensagens trocadas anteriormente, ou é capaz de responder a pedidos mais longos, o que poderá ser considerado como um serviço de qualidade superior, na perspetiva de um utilizador. Por seu turno, na produção de imagens, a resolução das imagens produzidas é uma escolha do programador. Em ambos exemplos, um modelo com mais memória ou maior resolução de imagem será superior, mas exigirá maior capacidade de computação.

IV.2. Barreiras à entrada e à expansão no fornecimento de modelos-base

O desenvolvimento e a implementação de modelos-base parecem estar sujeitos a fortes efeitos de escala e a custos de mudança, resultando em barreiras à entrada e à expansão no fornecimento de modelos-base. Estas barreiras são passíveis de criar uma tendência para elevada concentração nos mercados, beneficiando incumbentes e as primeiras empresas a desenvolver modelos de IA Generativa.

A estrutura de custos associada ao desenvolvimento e treino de modelos de IA Generativa é propícia à existência de economias de escala. Destaca-se, em particular, os custos iniciais associados à capacidade de computação, ao *know-how*, à eventual aquisição de dados e ao pré-processamento de grandes volumes de dados. À medida que se desenvolvem e se treinam modelos, o custo por modelo tenderá a reduzir-se, e, possivelmente, alcançando um desempenho superior (*vide* secções III.1 e III.2).

Os modelos-base permitem ainda gerar economias de gama na medida em que podem ser adaptados a um grande número de tarefas e ser utilizados para uma ampla gama de serviços. De facto, um modelo-base adaptado por *fine-tuning* tem geralmente melhor desempenho que um modelo treinado de raiz com os mesmos dados. A capacidade de

⁷⁷ E.g., o CLIP Score. Cf. Hessel et al. (2022). CLIPScore: A Reference-free Evaluation Metric for Image Captioning. Disponível [aqui](#).

⁷⁸ E.g. o FID (Fréchet Inception Distance) é um teste que permite avaliar a semelhança entre duas bases de dados de imagens. Cf. OpenAI (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents, disponível [aqui](#); ou Cf. Rombach et al. (2021). High-Resolution Image Synthesis with Latent Diffusion Models, disponível [aqui](#).

generalização dos modelos-base introduz, por este motivo, economias de gama entre modelos de IA Generativa a jusante, uma vez que os seus custos de desenvolvimento são partilhados. Quanto mais importante a capacidade de generalização dos modelos-base, maior a tendência para concentração em mercados de modelos-base, cada um sustentando um número significativo de modelos de *transfer learning*.

A implementação em massa de serviços de IA Generativa permite observar o comportamento e obter *feedback* dos utilizadores, podendo gerar uma relação positiva entre a dimensão e a qualidade dos modelos. Uma maior capacidade de recolha de dados de utilização, combinada com testes A/B, permite aos maiores fornecedores de IA Generativa intensificar a sua experimentação (cf. secção III.1). Por sua vez, essa maior experimentação permite treinar futuros modelos de forma mais eficiente e melhorar o seu desempenho, bem como a qualidade dos modelos e dos serviços a jusante. Nessa medida, pode gerar-se um círculo virtuoso entre o número de utilizadores e o desempenho dos modelos, i.e., **efeitos de rede**.

A observação do comportamento e a recolha de *feedback* de utilizadores pode permitir também a personalização dos modelos às suas preferências, o que pode gerar custos de mudança e um efeito de *lock-in*. Estes custos conferem uma vantagem competitiva aos fornecedores de maior dimensão e que já estejam estabelecidos no mercado, reforçando uma eventual tendência para a concentração no mercado.

No que respeita aos modelos de *transfer learning*, os efeitos de escala poderão não ser tão significativos. As técnicas de *transfer learning* são menos exigentes computacionalmente, por tomarem um modelo-base pré-treinado como ponto de partida. Estes modelos são também, por natureza, mais especializados, limitando a extensão das economias de gama. Por fim, implicam um menor número de escolhas quanto à arquitetura e hiperparâmetros, uma vez que estes são em larga medida definidos no desenvolvimento do modelo-base. Ainda assim, modelos de *fine-tuning* podem exigir bases de dados de elevada qualidade, que podem ter custos de aquisição elevados. Adicionalmente, à medida que o *fine-tuning* se torna mais extensivo, i.e., com mais alterações do modelo-base, aproxima-se do que seria o treino de um modelo-base, e os efeitos de escala tornar-se-ão mais relevantes.

As condições de concorrência nos mercados de modelos-base podem ter impacto nos mercados de modelos de *transfer learning*, e, em último caso, nos serviços de IA Generativa para o utilizador final. Os modelos de *transfer learning* podem ser treinados por operadores diferentes ou pelos mesmos operadores que desenvolveram o modelo-base. Existe, desse modo, uma relação vertical, com os modelos-base a montante e os modelos de *transfer learning*, a jusante.

Por fim, a legislação existente e futura de IA Generativa (cf. secção III.4) é passível de gerar barreiras à entrada e à expansão, afetando nomeadamente pequenos operadores e entrantes, que poderão ter menor capacidade de fazer face aos requisitos legais. Com efeito, um estudo de 2021 encomendado pela Comissão Europeia para avaliar o impacto dos requisitos da proposta inicial de Regulamento IA, estima os encargos

regulamentares em cerca de 17% do custo de investimento de uma empresa em IA, com maior impacto nas empresas de menor dimensão.⁷⁹ Nota-se, todavia, que algumas das medidas previstas na versão atual da proposta de Regulamento IA, e.g., *sandboxes* regulatórias, poderão contribuir para mitigar o impacto adverso dos encargos regulatórios nas empresas de menor dimensão e em fase de arranque (cf. Caixa 4).

IV.3.A relação com serviços de computação na *cloud* e fornecedores de *hardware* especializado

As condições de concorrência no setor dos serviços da *cloud* podem ter impacto nas condições de concorrência nos mercados de modelos-base. Os prestadores de modelos-base que não tenham acesso a infraestrutura própria têm de recorrer a prestadores de serviços de *cloud*. Nessa medida, a concorrência no setor da *cloud* pode ser importante para assegurar condições de acesso competitivas a capacidade de computação.

Os principais prestadores de serviços de *cloud* – Amazon Web Services (AWS), Microsoft Azure e Google Cloud (cf. Caixa 5) – disponibilizam serviços de *cloud* para empresas suas concorrentes nos mercados a jusante de serviços de IA Generativa. A título ilustrativo, a Google disponibiliza os seus modelos-base via a sua plataforma do Google Cloud, a plataforma de IA da Amazon está interligada com a AWS e a Microsoft disponibiliza os seus modelos-base no Microsoft Azure.⁸⁰

Alguns prestadores de *cloud* e de modelos-base também têm estabelecido parcerias entre si, envolvendo inclusive condições de exclusividade. A título ilustrativo, em 2019, a Microsoft fez um investimento inicial de mil milhões de dólares na OpenAI que visava “desenvolver uma plataforma de *hardware* e *software* no Microsoft Azure dimensionada para [IA Generativa]”, com a Microsoft como “fornecedor exclusivo de *cloud*” da OpenAI.^{81,82} A Anthropic, uma empresa de IA, tem uma parceria com a Google com vista a utilizar a *cloud* e os clusters de GPU e TPU⁸³ do Google Cloud para treinar e implementar os seus sistemas de IA.⁸⁴ Também através de uma parceria, a Hugging Face utiliza a AWS como fornecedor

⁷⁹ Renda, A. et al. (2022). Study to support an impact assessment of regulatory requirements for artificial intelligence in Europe, para a Comissão Europeia, disponível [aqui](#).

⁸⁰ CMA (2023). “AI Foundation Models Initial Report”, paragraph 4.22, disponível [aqui](#).

⁸¹ Greg Brockman (OpenAI) (2019). “[Microsoft Invests In and Partners with OpenAI to Support Us Building Beneficial AGI](#)”, 22.07.2019, acedido a 17.10.2023.

⁸² Desde então, a Microsoft tem vindo a reinvestir na OpenAI, mantendo a Azure como fornecedor de *cloud* exclusivo da OpenAI e com vista a implementar os modelos da OpenAI nos produtos da Microsoft (e.g., GitHub Copilot, DALL-E 2 e ChatGPT). Microsoft Corporate Blogs (2023) “[Microsoft and OpenAI extend partnership](#)” 23.01.2023, acedido a 17.10.2023.

⁸³ I.e., *Tensor Processing Units* (TPU) são aceleradores de IA personalizados, projetados pelo Google Cloud, e com vista a otimizar o treino e a inferência de grandes modelos de IA. Cf. informação disponibilizada [aqui](#) pelo Google Cloud sobre os seus produtos, acedido a 22.10.2023.

⁸⁴ Anthropic (2023). “[Anthropic Partners with Google Cloud](#)” 03.02.2023, acedido a 17.10.2023.

privilegiado de computação na *cloud* de forma a treinar, *fine-tune* e implementar os seus modelos na *cloud* da Amazon.⁸⁵

De igual modo, as condições de concorrência no mercado dos GPU⁸⁶ podem ter impacto nas condições de concorrência nos mercados de IA Generativa. Com efeito, os GPU são essenciais para o treino e para a implementação de modelos de IA Generativa (cf. secção III.2). Todavia, o mercado de GPU e, em particular dos GPU específicos para IA Generativa, é concentrado, destacando-se a Nvidia.⁸⁷ Esta empresa fornece grande parte dos GPU especializados para os modelos atualmente disponíveis de IA Generativa. Para além de ser o principal fornecedor de GPU específico para IA Generativa, a NVIDIA também desenvolveu uma série de serviços de IA Generativa de texto e imagem (e.g., Nvidia NeMo e Picasso).⁸⁸

O crescente interesse nos serviços de IA Generativa acentuou a necessidade de GPU, contribuindo para uma escassez deste *input*.⁸⁹ Esta escassez amplifica as barreiras à entrada de novos operadores nos mercados de IA Generativa. No entanto, também tem contribuído para um maior interesse, incluindo por parte de fornecedores de modelos-base, em investir na produção própria de GPU para IA Generativa ou através de parcerias, intensificando, por sua vez, as relações verticais ao longo da cadeia de valor.⁹⁰

As relações verticais podem trazer benefícios em termos de eficiências e permitir empresas de menor dimensão de competirem sem deterem recursos significativos, e.g., ao nível de computação. No entanto, também podem proporcionar riscos de concorrência (cf. secção VI). Em particular, a integração vertical é passível de conferir a capacidade e, em determinadas circunstâncias, o incentivo às empresas em excluir rivais nos mercados a jusante, nomeadamente através da degradação das condições de acesso aos modelos-base ou a serviços a montante necessários à sua implementação.⁹¹

⁸⁵ Hugging Face Blog (2023). "[Hugging Face and AWS partner to make AI more accessible](#)", acessido a 22.10.2023.

⁸⁶ Em 09.2023, a AdIC efetuou inspeções nas instalações de uma empresa suspeita de práticas anticoncorrenciais no setor dos GPU. AdIC. (2023) "[The General Rapporteur of the Autorité de la concurrence indicates that an unannounced inspection was carried out in the graphics cards sector](#)".

⁸⁷ CMA (2023). "[AI Foundation Models Initial Report](#)", para. 3.21; e Vipra & West (2023) "Computational Power and AI", AI Now Institute.

⁸⁸ Financial Times (2023). "[How Nvidia created the chip powering the generative AI boom](#)", 26.05.2023, acessido a 17.10.2023; NVIDIA Blog: <https://blogs.nvidia.com/blog/2023/03/21/gtc-keynote-spring-2023/>, 21.03.2023, acessido a 17.10.2023.

⁸⁹ Amba Kak and Sarah Myers West (2023). "[Computational Power and AI](#)", AI Now. A Microsoft também identifica, no seu relatório anual e contas, a escassez de GPU como um risco operacional ([Relatório Anual da Microsoft Corporation, Form 10-K](#), de 06.2023).

⁹⁰ Reuters News (2023). "[ChatGPT-owner OpenAI is exploring making its own AI chips](#)", acessido a 22.10.2023.

⁹¹ CMA (2023). "AI Foundation Models Initial Report", paragraphs 4.48-4.52, disponível [aqui](#).

Caixa 5 – Concorrência no setor da *cloud*

A utilização crescente de serviços de infraestruturas na *cloud* tem suscitado estudos sobre a concentração no mercado e os seus efeitos a jusante.

A oferta de serviços de infraestruturas na *cloud* é concentrada. Em 2023, a AWS e a Microsoft Azure eram os dois principais fornecedores destes serviços a nível mundial, sendo a Google o seu concorrente mais próximo. No Reino Unido, a Ofcom estima que a AWS e a Microsoft tinham, em 2022, uma quota de mercado combinada de 70% a 80% e que a Google tinha uma quota de 5% a 10%.⁹²

As três maiores empresas são conhecidas como “*hyperscalers*”, com uma larga rede de centros de dados e uma vasta gama de serviços, beneficiando de economias de escala. Existem fornecedores mais pequenos com infraestruturas; fornecedores independentes de software (ISV) que fornecem serviços de computação na *cloud*, mas não possuem uma infraestruturas subjacente; e outros fornecedores, como revendedores, consultores, etc.⁹³

A concorrência entre fornecedores de serviços de *cloud* centra-se principalmente na captação de novos clientes, quando estes escolhem pela primeira vez um fornecedor. Este primeiro momento traz benefícios para os clientes, incluindo a inovação de produtos e uma vasta escolha de serviços de software de ISV.⁹⁴

Um conjunto de autoridades⁹⁵ já identificaram barreiras à mudança e à utilização de múltiplas *cloud* após a escolha inicial de um fornecedor, o que pode limitar a concorrência no mercado. Foram destacadas as seguintes preocupações: (i) estrutura tarifária complexa e descontos utilizados para incentivar os clientes a recorrerem a um único fornecedor de serviços de computação na *cloud*, (ii) taxas “*egress*” (i.e., custos de transferência de dados da *cloud* para outro fornecedor) que induzem um efeito de *lock-in* e (iii) baixo nível de interoperabilidade e de portabilidade dos dados.

Foi também identificado, pela AdIC e pela ACM, que é provável que os regulamentos em discussão, como o Regulamento dos Dados,⁹⁶ possam vir a assegurar uma maior portabilidade e interoperabilidade dos dados neste setor.

V. CONCORRÊNCIA A JUSANTE DOS MODELOS-BASE DE IA GENERATIVA

Uma vez treinados, os modelos-base são adaptados a necessidades específicas, inclusive por operadores distintos a jusante da cadeia de valor. Em particular, os modelos-base são sujeitos a técnicas de *transfer learning*, nomeadamente *fine-tuning*, e

⁹² Ofcom (2023). “Cloud services market study – Final report”, p. 3, disponível [aqui](#).

⁹³ Idem, pp. 39-41.

⁹⁴ Idem, p. 3; e ACM (2022). “Market Study Cloud services” Case no. ACM/21/050317, disponível [aqui](#).

⁹⁵ AdIC (2023). “[Summary of Opinion 23-A-08 of 29 June 2023 on competition in the cloud sector](#)”; ACM (2022). “[Market Study Cloud services](#)”; Ofcom (2023). “[Cloud services market study – Final report](#)”. Entretanto, outras autoridades abriram estudos neste setor. Em junho de 2023, a FTC lançou uma consulta pública, sendo que os comentários se encontram disponíveis [aqui](#). Em outubro de 2023, a CMA publicou um [issues statement](#) com a identificação das teorias de dano a serem analisadas na investigação de mercado sobre os serviços de *cloud*.

customizados (e.g., via *plugins*). Estes passos são passíveis de ser desenvolvidos por operadores distintos, estabelecendo-se relações verticais e horizontais entre operadores relevantes do ponto de vista da concorrência (cf. Figura 4).

V.1. Modelos de *fine-tuning* de IA Generativa

O acesso a modelos-base tem um papel central na dinâmica concorrencial dos mercados de IA Generativa a jusante. O tipo de modelo-base, a sua dimensão, assim como a velocidade e o ritmo de acesso⁹⁷ ao modelo-base são características determinantes na capacidade dos prestadores de serviços de IA Generativa a jusante concorrerem entre si.

Há várias formas de conceber o acesso a modelos-base, mais ou menos abertas. Os formatos de acesso a modelos-base mais abertos são em regime de código aberto, mitigando a dependência de terceiros a modelos-base pagos. Neste caso, os ficheiros que contêm os parâmetros do modelo são partilhados, tal que um terceiro pode transferi-los e correr o modelo-base num ambiente controlado por si, seja localmente ou em *cloud*.⁹⁸ O acesso em código aberto pode exigir maior sofisticação por parte de terceiros, mas confere-lhes maior flexibilidade quanto ao tipo de treino que podem fazer e aos modelos que podem desenvolver.

Dentro do regime em código aberto pode haver diferenças quanto ao grau de abertura, restringindo até a possibilidade de aplicações comerciais, o que limitará a oferta nos serviços de IA a jusante. Por exemplo, um regime mais aberto pode partilhar também o código-fonte do modelo, documentação detalhada sobre o código-fonte e o modelo, dados de treino ou dados de monitorização do modelo. Também pode haver diferenças quanto aos direitos de acesso e de utilização do modelo-base.⁹⁹

Até à data, o acesso a modelos pré-treinados em código aberto tem sido relativamente comum no setor, o que terá contribuído positivamente para os avanços na IA Generativa. Existem, inclusive, inúmeras páginas na Internet que disponibilizam gratuitamente modelos pré-treinados, incluindo módulos de bibliotecas de programação,¹⁰⁰

⁹⁶ À data da redação, o processo legislativo relativo ao Regulamento dos Dados ainda se encontrava em curso. Este Regulamento visa complementar o [Regulamento Governação dos Dados](#). Comunicado, Comissão Europeia (2023). “[Data Act: Commission welcomes political agreement on rules for a fair and innovative data economy](#)”, acessado a 20.10.2023.

⁹⁷ E.g., quantas solicitações podem ser feitas por minuto ou quantidade de informação na solicitação ou na resposta.

⁹⁸ E.g., O LLAMA da Meta é um modelo disponibilizado em código aberto, que tem gerado inúmeras variantes, nomeadamente pela continuação do treino do modelo-base ou através de *fine-tuning*. Cf. Zhao et al. (2023). *A Survey of Large Language Models*, pp. 10-11.

⁹⁹ Por exemplo, o LLAMA da Meta tem limitações, nomeadamente, ao nível da possibilidade de uso comercial. CMA (2023). “*AI Foundation Models Initial Report*”, paragraph 3.105 e respetivas citações, disponível [aqui](#).

¹⁰⁰ E.g., o repositório do TensorFlow, uma biblioteca de desenvolvimento de IA e *machine learning* na linguagem de programação Python. Disponível [aqui](#).

repositórios dedicados de modelos¹⁰¹ ou repositórios de publicações académicas de IA que incluem recursos para replicação.¹⁰²

Em formatos de acesso mais fechados, terceiros não têm acesso direto ao modelo, fazendo-o sobretudo por via de API. Neste formato, um terceiro faz *upload* de dados de treino num servidor do fornecedor do modelo-base. Posteriormente, realiza-se o treino de *fine-tuning* e é criado o modelo de *fine-tuning*. O terceiro pode então fazer pedidos ao novo modelo, que responde através da API.¹⁰³

O acesso a modelos pré-treinados por API está intimamente relacionado com serviços de computação na *cloud*. O treino de *fine-tuning* exige capacidade de computação. Se os terceiros não têm acesso direto ao modelo, este será treinado numa *cloud* detida pelo fornecedor do modelo pré-treinado ou em parceria com este.

Também podem vir a desenvolver-se formatos de acesso a modelos pré-treinados, a partir de plataformas e aplicações dedicadas para submeter dados de treino de *fine-tuning*¹⁰⁴, e utilizar os modelos resultantes, de modo análogo ao que existe na publicidade digital para melhorar a *targeting* de anúncios.¹⁰⁵

V.2. Plugins de IA Generativa

Os *plugins* são componentes de *software* que podem ser usados para expandir as funcionalidades de um modelo de IA Generativa. Genericamente, os *plugins* conectam a IA Generativa a uma API de um serviço e fazem solicitações. O serviço com o qual a IA Generativa conecta pode ser uma API normal (i.e., não “inteligente”) ou outra IA Generativa. Conectar uma IA Generativa a outra pode expandir significativamente as suas funcionalidades. Por exemplo, uma IA Generativa de texto pode, a pedido do utilizador, fazer solicitações a uma IA Generativa de imagem. Assim, do ponto de vista do utilizador, a IA Generativa de texto produz texto e imagem.¹⁰⁶

Através dos *plugins*, a IA Generativa consegue fazer uso de mais informação que a sintetizada durante o seu treino (e.g., informação atualizada em tempo real). Os *plugins* podem também expandir o tipo de conteúdos que a IA Generativa pode produzir, ou o tipo de análises e de processamento de informação que a IA pode fazer.

¹⁰¹ E.g., o repositório Hugging Face, dedicado para a partilha de modelos e bases de dados para IA e *machine learning*. Disponível [aqui](#).

¹⁰² E.g., o repositório PapersWithCode, que inclui algumas das publicações académicas mais recentes em IA e *machine learning* juntamente com o código, base de dados e modelos para replicar essas publicações. Disponível [aqui](#).

¹⁰³ E.g., [API de acesso ao GPT-3.5](#).

¹⁰⁴ Um exemplo de plataforma dedicada é o [AutoTrain da Hugging Face](#).

¹⁰⁵ Veja-se, por exemplo, [o serviço de segmentação por lista de clientes, disponível no Google Ads](#).

¹⁰⁶ E.g., a modalidade Plus e Enterprise do ChatGPT – versões pagas do ChatGPT – introduziu em outubro de 2023 um *plugin* que permite ao ChatGPT fazer pedidos ao Dall-E 3. Cf. OpenAI (2023). Dall-E 3 in ChatGPT, disponível [aqui](#).

O que é que os *plugins* podem vir a trazer para os utilizadores?

Apresentam-se alguns exemplos do potencial dos *plugins*, no caso de IA Generativa de texto:



Análise financeira

Um *plugin* de dados financeiros permite à IA aceder a dados financeiros em tempo real. A partir dessa informação, a IA é capaz de fazer análises financeiras de uma empresa de forma automática.



Planeamento de viagens

Num *plugin* de planeamento de viagens, um utilizador reporta à IA para onde pretende viajar, em que dias e o que pretende fazer na viagem. Com base nessa informação, a IA apresenta possíveis estadias e voos. Quando o utilizador escolher uma, a IA envia um *link* para a página da agência de viagem, para o utilizador fazer a reserva.



Recomendação de produtos

Um utilizador que procure um determinado produto pode descrever à IA as suas características. Um *plugin* de recomendação de produtos pode conectar-se com lojas *online* ou com páginas de avaliações de produtos e encontrar o que o utilizador procura. Também pode enviar ao utilizador uma hiperligação para comprar o produto *online*.



Motor de busca e Internet

A IA pode conectar-se a um motor de busca para pesquisar informação que responda ao pedido do utilizador. Depois de visitar várias páginas nos resultados de pesquisa, compila as respostas num resumo que envia ao utilizador.



Questões sobre documentos e resumo de documentos

Neste *plugin*, o utilizador envia documentos (e.g., PDF) à IA e pode fazer perguntas à IA sobre os conteúdos destes documentos. Este serviço permite, por exemplo, responder questões sobre o conteúdo de livros ou resumir capítulos.





Executar código de programação a pedido do utilizador

Uma das funcionalidades da IA Generativa de texto é gerar código de programação a partir de um pedido em texto. Com um *plugin* de interpretador de programação, esse código pode ser executado pela IA a pedido do utilizador.

Caixa 6 – Ilustração da *customer journey* no *plugin* Expedia no ChatGPT

1. Utilizador solicita plano de viagem

 I want to visit Lisbon. Help me plan my trip to Lisbon.

 Of course! Let's plan an exciting trip to Lisbon for you. We'll go step-by-step:

1. **Lodging:** First, we'll find a comfortable place for you to stay.
2. **Flights:** Next, we'll look for the best flights to get you to Lisbon.
3. **Activities:** We'll then explore some fun activities you can do while in Lisbon.
4. **Car Rentals:** Lastly, if you're interested, we can also find a car for you to rent.


Let's start with lodging.

Lodging in Lisbon:
To find the best accommodations for you, could you please provide me with the following information:

- How many people will be traveling with you?
- How many bedrooms would you prefer?
- Any specific amenities or features you'd like in your accommodation? (e.g., gym, pool, breakfast included, etc.)


Feel free to provide as much or as little detail as you'd like, and I'll do my best to find the perfect place for you!

2. Utilizador escolhe alojamento


 Used Expedia

Great choice! Here are some top accommodations in Lisbon that have a pool and are suitable for a solo traveler:

1. **Hotel A**
 - Description: This luxury hotel offers a full-service spa, a restaurant, and a 24-hour fitness center. It's a smoke-free property with free WiFi in public areas. Additionally, there's a bar/lounge, a snack bar/deli, and a sauna.
 - Location: Located in Lisbon City Center, it's just steps from Coliseu dos Recreios and Avenida da Liberdade. Rossio Square and Santa Justa Elevator are also within 2 mi (3 km).
 - Star Rating: ★★★★★
 - Guest Rating: 9.2/10
 - Average Nightly Price: \$296.03 (with a 10% promotion)
 - Check-in: 5th January 2024
 - Check-out: 9th January 2024

 Hotel A is fine.


3. A IA disponibiliza *link* para compra e oferece outros serviços


 Fantastic choice! **Hotel A** is a luxurious and elegant hotel located in the heart of Lisbon's historic district. With its full-service spa, top-notch amenities, and close proximity to major attractions, you're sure to have a memorable stay.

You can book your stay at **Hotel A** [here](#).

Now, let's move on to the next step. Would you like assistance with finding the best flights to Lisbon? If yes, could you please provide me with your departure city or airport? If not, we can jump to activities or car rentals. Let me know how you'd like to proceed! 😊


4. Utilizador escolhe voo


 I will be departing from Madrid.

 Great! Let's find the best flights for you from Madrid to Lisbon Humberto Delgado Airport.

Here are the top flight recommendations for your journey:

1. **Flight A**
 - Departure: 5th January 2024, 10:50 PM from Madrid (MAD)
 - Arrival: 5th January 2024, 11:10 PM at Lisbon (LIS)
 - Duration: 1h 20m
 - Price: \$30.43
 - Non-stop flight

 Flight A is fine.

 Excellent choice! **Flight A** is a convenient and affordable option for your journey from Madrid to Lisbon.

You can book your flight on Expedia [here](#).

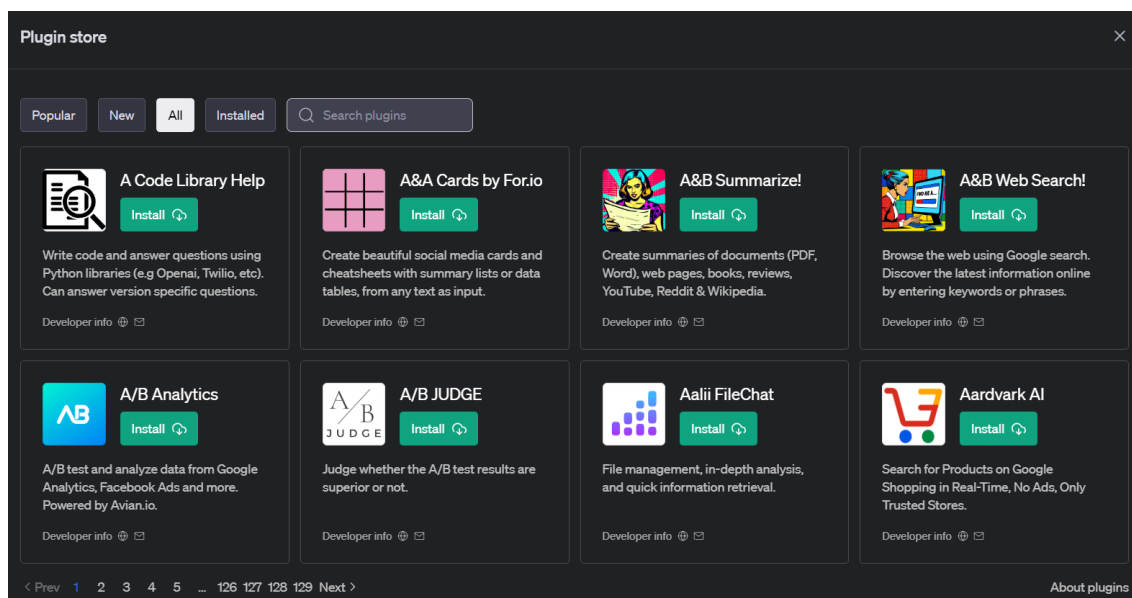
Now that we've sorted out your flight, would you like recommendations for activities to do in Lisbon? Or perhaps you're interested in car rentals? Let me know how you'd like to proceed, and I'll assist you further! 😊

Fonte: Pedido feito no *plugin* Expedia, no ChatGPT Plus.

Os *plugins* podem ser desenvolvidos por terceiros, permitindo expandir o número de fornecedores dedicados à IA Generativa. A título ilustrativo, a OpenAI anunciou, em março

de 2023, que disponibilizou a programadores terceiros de *plugins*, a documentação necessária para criar um *plugin* para o ChatGPT. Os primeiros *plugins* foram criados por Expedia, FiscalNote, Instacart, KAYAK, Klarna, Milo, OpenTable, Shopify, Slack, Speak, Wolfram e Zapier.¹⁰⁷ Atualmente, existem mais de 1000 *plugins* disponíveis no ChatGPT.

Figura 5 – A *plugin store* no ChatGPT Plus



Fonte: <https://chat.openai.com/>, acessado a 26.10.2023.

Os *plugins* são uma possível fonte de efeitos de rede e um fator de concentração em mercados de IA Generativa. Os *plugins* permitem que se formem novos mercados que juntam, de um lado, utilizadores de *plugins* e, do outro, fornecedores de *plugins*. A IA Generativa que concentrar um maior número de *plugins* poderá ter um maior leque de funcionalidades e, por isso, ser mais atrativa para utilizadores finais. Assim, na medida em que os *plugins* se tornarem um elemento estruturante da oferta de serviços de IA Generativa, é possível que se desenvolvam no setor ecossistemas em torno dos principais fornecedores, de modo a capitalizar os efeitos de rede. A organização da IA Generativa em ecossistemas, pautados por efeitos de rede, pode significar uma concentração elevada nestes mercados.

Os *plugins* são, assim, análogos a aplicações para sistemas operativos, devido à sua capacidade de expandir as funcionalidades da IA Generativa, à sua forma de distribuição e às preocupações de concorrência associadas.

Os *plugins* podem ser distribuídos de diversos modos, conforme o serviço da IA Generativa. Em serviços mais orientados para utilizadores finais, os *plugins* podem ser distribuídos via *plugin stores*.¹⁰⁸ Analogamente ao que sucede em *app stores*, os utilizadores

¹⁰⁷ Cf. OpenAI Blog (2023). “[ChatGPT plugins](#)”, acessado a 22.10.2023.

¹⁰⁸ A título ilustrativo, o ChatGPT Plus permite o acesso a *plugins*, de forma análoga a uma *app store*. Cf. Documentação da OpenAI, disponível [aqui](#), e <https://chatgpt.plugin.support/>, acessados a 22.10.2023.

podem pesquisar e instalar *plugins* na sua conta (i.e., interface do utilizador) do serviço de IA Generativa.¹⁰⁹ Em serviços orientados para empresas ou utilizadores mais sofisticados, os *plugins* podem ser distribuídos de outro modo, conforme o grau de abertura do serviço de IA Generativa e a sofisticação dos utilizadores (e.g., em código aberto).

Quanto ao modelo de negócio dos *plugins*, parece ser ainda demasiado cedo para mapear que formas de monetização serão mais comuns no futuro.

VI. POSSÍVEIS RISCOS DE CONCORRÊNCIA

Muitos produtos e serviços digitais, incluindo a IA Generativa, apresentam características que os tornam suscetíveis a elevados níveis de concentração no mercado. O mesmo sucede em alguns mercados adjacentes, como sejam as infraestruturas de rede e de computação que sustentam o setor digital.

Nessa medida, alguns operadores podem obter um poder de mercado significativo, e ter a capacidade e o incentivo para introduzir e explorar possíveis estrangulamentos no mercado. Estes comportamentos, com vista a conferir uma vantagem competitiva a determinados produtos em detrimento de outros, são passíveis de desvirtuar o processo concorrencial e de prejudicar os consumidores.

Estes possíveis estrangulamentos no mercado baseiam-se, genericamente, no acesso e controlo de algum recurso necessário para colocar uma oferta competitiva no mercado. Exemplos incluem bases de dados, a capacidade de recolha de dados, a capacidade de computação, bases de utilizadores e efeitos de rede, ou a experiência acumulada do operador.

Esta secção visa apenas tipificar alguns possíveis riscos de concorrência que, dadas as características *supra* identificadas, podem ser, conceptualmente, enquadrados nos mercados em análise. Aborda-se ainda, de forma complementar, alguma da prática decisória da Comissão Europeia no setor digital.

VI.1. Riscos de alavancagem na integração de IA Generativa em outros produtos

Os fornecedores de serviços de IA Generativa são passíveis de ter incentivos para os integrar em outros dos seus produtos, numa lógica de ecossistema de produtos e serviços.¹¹⁰ Por exemplo, um serviço de *chatbot* pode ser integrado, ou fornecido conjuntamente, em motores de busca, sistemas operativos, *software* de produtividade ou serviços de computação na *cloud*.

¹⁰⁹ Note-se que as obrigações constantes do Regulamento dos Mercados Digitais poderão vir a ser aplicáveis aos *plugins* por se enquadrarem em serviços de intermediação em linha. Cf. Regulamento (UE) 2022/1925 do Parlamento Europeu e do Conselho de 14 de setembro de 2022 relativo à disputabilidade e equidade dos mercados no setor digital e que altera as Diretivas (UE) 2019/1937 e (UE) 2020/1828 (Regulamento dos Mercados Digitais), disponível [aqui](#).

¹¹⁰ AdC (2019). Issues Paper “Ecossistemas Digitais, Big Data e Algoritmos”, pp. 14-15, disponível [aqui](#).

Este tipo de integração, em lógica de ecossistema, é passível de resultar em melhores produtos e ofertas mais atrativas para os consumidores, nomeadamente pela conveniência de um único ponto de acesso.

No entanto, é também passível de se traduzir em riscos para a concorrência. Hipoteticamente, uma prática de subordinação (i.e., *tying*¹¹¹) de IA Generativa com produtos *core* pode servir para acelerar o seu desenvolvimento, em detrimento de IA concorrentes.¹¹² Um operador pode, por exemplo, alavancar o número muito elevado de utilizadores num produto/serviço *core*, canalizando-os para a sua IA Generativa. Se a IA Generativa beneficiar de efeitos de escala significativos, este comportamento pode conferir uma vantagem competitiva à IA Generativa do operador. Uma imposição de aquisição de um determinado produto/serviço para aceder a um modelo e/ou serviço de IA Generativa também será passível de ter um efeito de encerramento, em particular, se a empresa tiver posição dominante no mercado associado à IA Generativa. Os incentivos de encerramento de mercado irão depender, todavia, dos modelos de negócio e estratégias de monetização que venham a surgir quer relativamente aos modelos-base de IA Generativa, quer aos serviços a jusante, para além de outros fatores.

Em termos conceptuais, também podem surgir riscos para a concorrência caso um operador com poder de mercado em um determinado mercado de IA Generativa utilize a sua posição para favorecer outros dos seus produtos e serviços, integrando neles a sua IA Generativa de forma exclusiva. O poder de mercado do operador em IA Generativa pode resultar da acumulação de vantagens várias, como uma melhor otimização dos modelos ou o acesso a melhores bases de dados.

Assim, estratégias de alavancagem de um mercado para um outro serão tão mais preocupantes se levadas a cabo por uma empresa com posição dominante em determinado mercado (a montante, a jusante ou em mercados adjacentes) com vista a excluir concorrentes. Conforme se ilustra na Caixa 7, práticas como seja *tying* conduzidas por empresas com posição dominante são passíveis de violar o direito da concorrência.

¹¹¹ Segundo as Orientações de abusos de exclusão da Comissão, “[a] «subordinação» refere-se geralmente a situações em que os clientes que compram um produto (produto subordinante) são obrigados a comprar igualmente outro produto da empresa dominante (produto subordinado)”. Comunicação da Comissão - Orientação sobre as prioridades da Comissão na aplicação do artigo 82.º [102.º] do Tratado CE a comportamentos de exclusão abusivos por parte de empresas em posição dominante (p. 48).

¹¹² Este risco foi também identificado pela FTC e pela CMA, bem como em artigos. FTC Blog (2023) “[Generative AI Raises Competition Concerns](#)”; CMA (2023) “AI Foundation Models Initial Report”, para 4.27, disponível [aqui](#); Carugati, C. (2023). “Competition in generative artificial intelligence foundation models”, disponível [aqui](#).

Caixa 7 – Casos de abuso de posição dominante no setor digital, na UE

Destacam-se, de seguida, algumas investigações e decisões de abusos de posição dominante no setor digital, na UE.

Em julho de 2018, a CE sancionou a Google em 4,3 mil milhões de euros, por abuso de posição dominante nos mercados de serviços de pesquisas genéricas na Internet, nos sistemas operativos para *smartphones* e nas *app stores* para Android. Em particular, a CE identificou que a Google exigia aos fabricantes de *smartphones* que pré-instalasse o Google Search e o Google Chrome nos seus dispositivos Android como condição para aceder ao Google Play Store. A Google também implementou um esquema de incentivos financeiros para induzir os fabricantes de dispositivos e os operadores de rede a incluir o Google Search pré-instalado de forma exclusiva nos dispositivos móveis. A Google visava reforçar a sua posição no mercado de serviços de pesquisas genéricas.¹¹³

Em março de 2004, a CE emitiu uma decisão contra a Microsoft por abuso da sua posição dominante no mercado dos sistemas operativos.¹¹⁴ O abuso envolveu: (i) a recusa de fornecimento de informações de interoperabilidade indispensáveis para que os seus concorrentes pudessem competir no mercado dos sistemas operativos de servidores de grupos de trabalho; e (ii) *tying* do seu leitor multimédia ao seu sistema operativo. O comportamento da Microsoft permitiu-lhe reforçar uma posição dominante no mercado dos sistemas operativos de servidores de grupos de trabalho e fragilizou a concorrência no mercado dos leitores multimédia. A CE aplicou à Microsoft uma coima de cerca de 497 milhões de euros e as seguintes medidas corretivas: (i) a divulgação e interoperabilidade de especificações de interface às empresas interessadas; e (ii) a oferta de uma versão do sistema operativo Windows que não incluísse o Windows Media Player.

Em maio de 2022, a AdC abriu um processo de contraordenação visando a Google pela possível prática de um abuso de posição dominante na publicidade online.¹¹⁵ Em causa estiveram comportamentos da Google que indiciariam um possível favorecimento desta empresa em vários níveis da cadeia de valor associada à venda de espaço publicitário online. A investigação da AdC centrou-se na possibilidade de a Google ter utilizado informação sobre os leilões para anúncios online, à qual os seus concorrentes não tinham acesso, para condicionar, em seu favor, o resultado desses leilões. Em julho de 2022, a CE informou a AdC de que pretendia alargar o âmbito da sua própria investigação à Google para incluir as práticas e os mercados objeto de investigação pela AdC em Portugal. Em setembro de 2022, a AdC encerrou a investigação, passando a mesma a ser conduzida pela CE.

Em junho de 2023, a CE anunciou que enviou uma Nota de Ilícitude à Google, por ter alegadamente favorecido os seus próprios serviços de intermediação de publicidade *display*. Este setor junta, de um lado, anunciantes que querem exibir os seus anúncios a consumidores e, do outro lado, *publishers* que pretendem vender espaço publicitário. No setor, os anúncios são selecionados através de leilões. A CE identificou que a Google terá beneficiado os seus próprios serviços de intermediação publicitária, em condições desiguais, enquanto organizador e participante em leilões para selecionar os anúncios que serão exibidos num dado espaço publicitário. Adicionalmente, a Google terá alavancando o número de anunciantes que usam os seus serviços em favor dos seus outros serviços de intermediação publicitária, nomeadamente evitando fazer licitações em leilões organizados por empresas concorrentes à Google.¹¹⁶

Em julho de 2023, a CE anunciou a abertura de investigação visando a Microsoft pela possível prática de *tying* do seu produto Teams face aos seus produtos *core* para empresas Office 365 e Microsoft 365, com vista a defender a sua posição no mercado do *software* de produtividade.¹¹⁷

VI.2. Acesso privilegiado a bases de dados para o desenvolvimento de IA Generativa

Um operador que forneça serviços de IA Generativa pode ter acesso a bases de dados importantes e dificilmente replicáveis por concorrentes, por exemplo, fruto da sua presença noutros mercados digitais. A presença de um operador em vários mercados digitais em simultâneo é comum no setor digital dada a sua organização, em grande medida, em ecossistemas digitais que oferecem conjuntamente vários produtos e serviços.¹¹⁸ Por sua vez, as bases de dados, nomeadamente obtidas através de outros produtos, podem ser cruciais para o desenvolvimento de modelos de IA Generativa, sejam modelos-base ou modelos de *fine-tuning*.

Um operador pode ter o incentivo para conceder acesso privilegiado ou preferencial a estas bases de dados à sua própria IA Generativa, em detrimento de IA concorrentes.

Este comportamento é passível de distorcer a concorrência, em particular caso as bases de dados assumam um papel instrumental para o desenvolvimento de serviços de IA concorrentes e o operador com acesso privilegiado tenha uma posição dominante no mercado relevante.¹¹⁹ O impacto adverso nas condições de concorrência, decorrente de um acesso privilegiado a bases de dados, será tanto maior quanto maior a essencialidade desses dados e se a empresa em causa usufrui de poder de mercado significativo.

VI.3. Acesso privilegiado a modelos de IA Generativa

Os modelos-base são o pilar dos serviços de IA Generativa a jusante, mais próximos dos utilizadores finais. Os fornecedores de modelos-base estão, por esse motivo, numa posição privilegiada para desenvolver e implementar os seus próprios serviços de IA Generativa a jusante. Em particular, os fornecedores de modelo-base podem ter maior

¹¹³ Decisão da Comissão de 18.07.2018 relativa a um processo nos termos do artigo 102.º do TFUE e do artigo 54.º do Acordo EEE (Processo AT.40099 – Google Android), disponível [aqui](#).

¹¹⁴ Decisão da Comissão de 24.03.2004 relativa a um processo nos termos do artigo 82.º do TFUE (Processo COMP/C-3/37.792 Microsoft), disponível [aqui](#).

¹¹⁵ A ficha do processo está disponível [aqui](#). Vide, também, AdC (2022) Nota de Acompanhamento “Defesa da concorrência no setor digital em Portugal”.

¹¹⁶ Comunicado da Comissão Europeia (2023) “[Antitrust: Commission sends Statement of Objections to Google over abusive practices in online advertising technology](#)”, 14.06.2023.

¹¹⁷ Comunicado da Comissão Europeia (2023) “[Antitrust: Commission opens investigation into possible anticompetitive practices by Microsoft regarding Teams](#)”, 27.07.2023.

¹¹⁸ AdC (2019). Issues Paper “Ecossistemas Digitais, Big Data e Algoritmos”, pp. 8-12, disponível [aqui](#).

¹¹⁹ Não obstante não se ter verificado no setor digital, refira-se que, em 2015, a AdC condenou a Associação Nacional de Farmácias (ANF) pela prática de esmagamento de margens, por via do acesso discriminatório a dados comerciais das farmácias. No caso, estes dados foram considerados como *input* no mercado a jusante de estudos de mercado, onde participavam a ANF e outros concorrentes. Decisão da AdC PRC/2009/13, disponível [aqui](#).

facilidade em adaptá-los a tarefas específicas, beneficiando (i) de conhecimento especializado acerca do seu próprio modelo-base, (ii) da capacidade de computação que lhe permitiu desenvolver o modelo-base e (iii) da eventual capacidade de generalização do modelo-base para desempenhar um grande número de tarefas específicas.

Assim sendo, os fornecedores de modelos-base podem ter capacidade e incentivos para conferir um tratamento privilegiado ou preferencial aos seus próprios serviços de IA Generativa a jusante, em detrimento de outros serviços de IA Generativa a jusante. Este tratamento privilegiado pode materializar-se, por exemplo, pelo acesso exclusivo da IA Generativa do fornecedor de modelos-base à melhor versão do modelo-base (e.g., versão com mais parâmetros, treinada com bases de dados exclusivas, a mais recente). Em alternativa, por exemplo, o fornecedor de modelos-base pode introduzir limitações do ritmo de acesso¹²⁰ ao modelo-base a operadores concorrentes, mas não à sua própria IA Generativa.

Este risco será maior caso o fornecedor do modelo-base tenha poder de mercado, e o comportamento em causa seja passível de criar barreiras ao desenvolvimento e implementação de serviços de IA Generativa concorrentes. Nesse cenário, é possível que o comportamento limite, de forma significativa, a inovação de serviços de IA Generativa.

As mesmas questões podem aplicar-se a jusante, nomeadamente no acesso a modelos de *fine-tuning*. Um fornecedor de modelos de *fine-tuning* (ou de outros modelos derivados de modelos-base) estará numa posição privilegiada para desenvolver os seus próprios *plugins* e pode ter a capacidade e os incentivos para conferir um tratamento privilegiado ou preferencial aos seus próprios *plugins*. Adicionalmente, o fornecedor do modelo de *fine-tuning* poderá ter algum grau de controlo sob a forma como os *plugins* são distribuídos (e.g., *plugin stores*), o que introduz potenciais mecanismos adicionais de exclusão. A existência ou não de incentivos de encerramento de mercado irá, todavia, depender dos modelos de negócio e estratégias de monetização que venham a surgir, para além de outros fatores.

VI.4. Exigência de contrapartidas abusivas para o fornecimento de IA Generativa

Um fornecedor de IA Generativa com poder de mercado pode ter a capacidade e o incentivo para exigir contrapartidas abusivas a outros operadores ou utilizadores que pretendam aceder à sua IA Generativa. Para além dos custos de acesso, os termos e condições de acesso à IA Generativa podem incluir direitos do fornecedor de IA sobre os conteúdos produzidos e direitos sobre conteúdos, dados e outra informação submetida pelos próprios operadores ou utilizadores.

Em particular, o acesso a IA Generativa pode ser concedido na condição que o operador ou o utilizador cedem todos os conteúdos ou dados submetidos ao fornecedor do modelo de IA Generativa. Por exemplo, o *fine-tuning*, que adapta o modelo-base a uma tarefa específica, implica um treino adicional a partir do modelo-base. Este treino adicional requer dados adicionais, que o operador que acede ao modelo-base detém, mas

¹²⁰ Ver nota de rodapé 97.

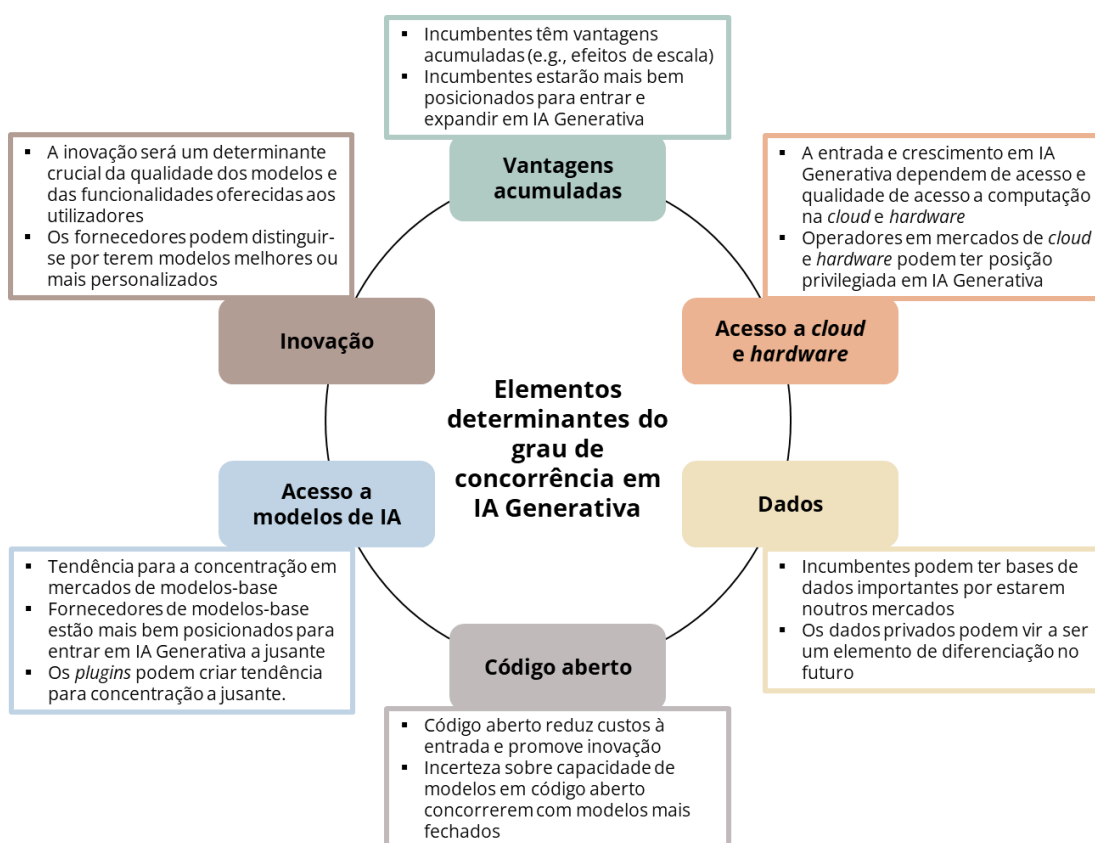
não o fornecedor do modelo-base. Ora, o fornecedor do modelo-base, como contrapartida de acesso, pode exigir o direito de utilizar esses dados de *fine-tuning* para o treino dos seus próprios modelos.

Em suma, de forma a assegurar que a economia venha a usufruir dos potenciais benefícios da IA Generativa, importa que não se verifiquem barreiras significativas que venham a impedir o aparecimento de serviços alternativos de IA Generativa a jusante. A este respeito, destaca-se que **as empresas com posição dominante têm a responsabilidade de garantir que as suas escolhas e comportamentos não distorcem indevidamente a concorrência.**

VII. MAPEAMENTO DOS DETERMINANTES DE CONCORRÊNCIA EM IA GENERATIVA

Considerando a experiência acumulada no setor digital por autoridades da concorrência e os riscos identificados *supra*, é **possível mapear alguns dos elementos determinantes do grau de concorrência no setor de IA Generativa.**

Figura 6 – Os elementos determinantes do grau de concorrência em IA Generativa



Fonte: AdC.

Vantagens acumuladas por incumbentes no setor digital

A evolução do setor de IA Generativa poderá ser largamente determinada pelas condições de concorrência atuais do setor digital como um todo. De modo geral, o setor digital caracteriza-se por fortes efeitos de escala que tornam muitos dos respetivos mercados suscetíveis a elevados níveis de concentração. Uma consequência destas características são os incentivos e vantagens no desenvolvimento de ecossistemas, resultando em vantagens competitivas às empresas já estabelecidas no mercado.

A IA Generativa surge num contexto em que os incumbentes digitais têm fortes vantagens acumuladas. O facto de haver incentivos e vantagens em organizar mercados digitais em ecossistemas pode significar que os incumbentes digitais estarão melhor posicionados para entrar e expandir em mercados de IA Generativa, reforçando a sua posição no setor digital.

Para aproveitar as oportunidades da IA Generativa para a economia, é importante que as autoridades da concorrência se mantenham vigilantes. Importará assegurar que o crescimento de operadores nos mercados de IA Generativa é feito através da concorrência pelo mérito e não pela alavancagem de poder de mercado.

Acesso a serviços de computação na cloud e hardware especializado

A entrada e o crescimento de operadores nos mercados de IA Generativa dependerá do acesso a capacidade de computação e da qualidade desse acesso. Os operadores têm duas vias para aceder a capacidade de computação: (i) serviços de computação na *cloud*; e (ii) construção de infraestrutura própria pela aquisição de *hardware* especializado, como CPU e GPU. Este acesso depende das empresas presentes nos mercados de computação na *cloud* ou *hardware* especializado que, por sua vez, poderão estar numa posição privilegiada para entrar e crescer em mercados de IA Generativa.

Assim, a promoção da concorrência nos mercados de IA Generativa estará também dependente da promoção da concorrência nestes mercados a montante. Mercados a montante mais competitivos tornarão mais competitivo o acesso aos *inputs* necessários para o desenvolvimento e implementação da IA Generativa, em benefício dos consumidores.

Bases de dados

Alguns incumbentes digitais podem ter vantagens acumuladas quanto a bases de dados, que constituem um dos elementos-chave de concorrência em IA Generativa. Quer o volume quer a qualidade dos dados de treino utilizados têm um impacto determinante no desempenho de serviços de IA Generativa e na sua capacidade de competir no mercado. O acesso a bases de dados já existentes por parte dos incumbentes digitais permitirá acumular vantagens, mas também a capacidade de gerar novos dados no futuro. Uma das principais fontes de vantagens em dados são grandes bases de utilizadores, que são monitorizados ou produzem conteúdos nesses serviços. Esta informação pode ser, posteriormente, convertida em dados de treino para modelos de IA Generativa.

É possível que os dados privados venham a ganhar importância no futuro e a ser um elemento-chave de diferenciação dos serviços de IA Generativa. Uma parte relevante dos modelos de IA Generativa atualmente disponíveis recorrem a bases de dados públicas, no entanto, tal poderá alterar-se após a atual difusão inicial dos serviços de IA Generativa.

O papel do código aberto na IA Generativa

Alguns dos modelos de IA Generativa têm sido disponibilizados em formato de código aberto, permitindo a empresas terceiras oferecerem serviços a jusante baseados nesses modelos. Este acesso em regime aberto permite a terceiros a flexibilidade para usar, adaptar e expandir modelos de IA Generativa, reduzindo custos de entrada e contribuindo para uma maior inovação nas diversas aplicações da tecnologia.

No entanto, não é evidente o papel promotor de concorrência do código aberto. Por um lado, este regime pode não fornecer incentivos suficientes para o desenvolvimento de modelos competitivos face a regimes fechados. Por outro lado, será certamente necessária alguma forma de monetização o que levanta dúvidas quanto à concretização e viabilidade de modelos em regimes abertos. Acresce que os modelos em código aberto podem ser disponibilizados mediante condições que limitam a sua utilização comercial por terceiros. Este regime também não exclui possíveis problemas concorrenciais, na medida em que um modelo em código aberto pode ser utilizado para reforçar o poder de mercado do seu fornecedor num mercado adjacente.

Acesso a modelos de inteligência artificial

A capacidade de generalização dos modelos-base pode criar uma tendência para a concentração em mercados de modelos-base, com impacto a jusante. Estes modelos são *input* para os fornecedores de serviços de IA Generativa, pelo que a concentração de modelos-base resultará num menor poder negocial dos fornecedores a jusante. Os fornecedores de modelos-base estão também numa posição privilegiada para entrar e crescer em mercados de serviços de IA Generativa a jusante, o que poderá ter impacto nas condições de acesso dos seus modelos a empresas rivais a jusante.

A jusante, os *plugins* podem gerar efeitos de rede e criar uma tendência para a concentração nos modelos de *transfer learning*. Por esse motivo, as mesmas preocupações e dinâmica aplicam-se ao acesso de *plugins* a modelos de *transfer learning*. Os fornecedores destes modelos também estão numa posição privilegiada para entrar e crescer em eventuais mercados de *plugins*.

Inovação

Um elemento-chave no setor de IA Generativa será a inovação. A inovação pode operar via novos produtos e serviços, ou melhorando os já existentes. Quanto à melhoria de produtos e serviços existentes, os fornecedores podem produzir estritamente melhores que os concorrentes (diferenciação vertical), ou adaptando modelos às necessidades e preferências específicas de utilizadores (diferenciação horizontal).

A capacidade de generalização da IA Generativa sugere um potencial significativo de inovação. A concorrência, enquanto promotor da inovação, é fundamental para concretizar este potencial de inovação.

Como tal, o foco da defesa da concorrência no contexto da Inteligência Artificial terá necessariamente de estar nos seguintes elementos-chave: (i) acesso aos dados, (ii) acesso à computação na *cloud* ou *hardware* especializado e (iii) acesso aos modelos-base de IA Generativa. A AdC, no âmbito das suas atribuições e no contexto da cooperação internacional, manter-se-á atenta aos desenvolvimentos da inteligência artificial, e não hesitará em intervir para garantir a promoção e a aplicação do direito da concorrência em Portugal, em benefício dos consumidores e sempre que os riscos identificados se materializem.

IMPORTÂNCIA DA CONCORRÊNCIA NA INTELIGÊNCIA ARTIFICIAL GENERATIVA

Como pode a concorrência assegurar mais benefícios para os consumidores e empresas

-  1. A Inteligência Artificial Generativa traz um **universo de novos produtos e serviços para os consumidores e empresas**, exibindo capacidades tipicamente consideradas exclusivas de seres humanos.
-  2. O potencial de aplicações da Inteligência Artificial Generativa é aparentemente inesgotável, podendo ter **impacto transversal e significativo** em todos os setores da economia.
-  3. O setor da Inteligência Artificial Generativa está **em crescimento e tem inúmeras oportunidades para inovação**.
-  4. A concorrência será um **instrumento essencial para concretizar em pleno o potencial da Inteligência Artificial Generativa** e as oportunidades para a inovação, como o motor dos incentivos para a inovação.
-  5. É importante **mapear os elementos determinantes que afetam o processo concorrencial** e antecipar os riscos para a concorrência no setor da Inteligência Artificial Generativa.
-  6. **As autoridades da concorrência têm acumulado, nos últimos anos, experiência significativa no setor digital**, identificando os determinantes que, em cada mercado, podem produzir resultados lesivos para os consumidores e para as empresas.
-  7. A Inteligência Artificial Generativa é **emblemática dos desafios que o setor digital coloca**, numa ótica de concorrência. Os mercados de Inteligência Artificial Generativa têm tendência para a concentração, o que pode conferir capacidade e incentivos a alguns operadores para desvirtuar o processo concorrencial.