Blog

**Desbloquear dados para impulsionar o comércio e a cultura europeus**

*Por Brad Smith, Vice Chair & President da Microsoft*

**Lisboa, 21 de julho de 2025** – A Europa é o lar de mais de 200 línguas e de um património cultural rico que se estende por milhares de anos, preservado em milhões de bens culturais que contam a história dos seus povos. Mas estas línguas são mais do que veículos de herança e história — sustentam tanto a cultura como o comércio, ao tornarem possível que as pessoas se conectem, criem e façam negócios.

Contudo, à medida que o mundo se digitaliza, grande parte da diversidade linguística e cultural da Europa corre o risco de ser deixada para trás. A maioria do conteúdo disponível na internet — a principal fonte de dados de treino para os atuais grandes modelos de linguagem (LLMs) — está em inglês. E grande parte desse conteúdo reflete uma perspetiva americana. A Comissão Europeia já alertou que a ambição do continente de digitalizar o seu vasto acervo cultural continua “significativamente fora de alcance”. Como os líderes europeus têm reconhecido, sem uma ação urgente, este desequilíbrio não é apenas uma preocupação cultural — é também uma questão comercial. Uma inteligência artificial (IA) que não compreende as línguas, histórias e valores da Europa não pode servir plenamente os seus cidadãos, as suas empresas ou o seu futuro.

É por isso que hoje, em Paris, estamos a reforçar o nosso compromisso com o futuro digital da Europa com duas novas iniciativas centradas em tornar mais acessível e aberto aquilo que é exclusivamente europeu — as suas línguas e cultura. Isto dá continuidade aos nossos [Compromissos Digitais Europeus](https://blogs.microsoft.com/on-the-issues/2025/04/30/european-digital-commitments/), anunciados no início deste ano, que visam expandir a infraestrutura de IA e cloud, reforçar a resiliência digital e a proteção da privacidade de dados, melhorar a cibersegurança e apoiar a soberania digital da Europa e a sua economia em geral.

Em primeiro lugar, para apoiar o desenvolvimento de LLMs mais multilingues na Europa e para a Europa, vamos destacar colaboradores de dois dos nossos centros de inovação em Estrasburgo, França — uma cidade há muito tempo ponto de encontro de culturas e atualmente sede de instituições europeias fundamentais. Estes centros ajudarão a expandir a disponibilidade de dados multilingues para o desenvolvimento de IA — aproveitando o Microsoft Azure, a nossa experiência técnica e parcerias em toda a Europa para promover uma representação linguística mais inclusiva nos modelos de IA. Como parte deste esforço, estamos também a lançar um convite à apresentação de propostas para ajudar a aumentar a oferta de conteúdo digital em 10 línguas europeias.

Em segundo lugar, para garantir que a riqueza cultural da Europa está representada e acessível no mundo digital, estamos a expandir a iniciativa Culture AI da Microsoft, que ajuda a salvaguardar línguas, monumentos e artefactos através de réplicas digitais e colaboração em dados. Desde 2019, a Microsoft tem preservado digitalmente patrimónios como a Antiga Olímpia, na Grécia, o Monte Saint-Michel, em França, a Basílica de São Pedro, em Roma, e o 80.º aniversário dos Desembarques Aliados na Normandia, entre outros. Hoje, anunciamos que, neste outono, a Microsoft irá iniciar trabalhos com o Ministério da Cultura francês e a empresa francesa Iconem para criar uma réplica digital de Notre-Dame — a recém-restaurada obra-prima gótica de Paris, com 862 anos.

Este tipo de apoio à Europa e à sua diversidade não é novidade para a Microsoft. Estes passos mais recentes para apoiar as línguas e a cultura baseiam-se em mais de 40 anos de experiência a servir países e culturas por toda a Europa e pelo mundo. Desde cedo aprendemos que capacitar cada pessoa no planeta exige que as tecnologias que oferecemos estejam disponíveis nas línguas que o mundo fala. É por isso que, atualmente, o Windows suporta mais de 90 línguas, incluindo todas as línguas oficiais europeias e outras línguas regionais como o basco, o catalão, o galego, o luxemburguês, o valenciano, entre outras. O Microsoft 365 também tem um alcance alargado, com suporte nas aplicações do Office em mais de 30 línguas europeias, incluindo todas as línguas oficiais da União Europeia (UE).

**A urgência de colmatar a lacuna linguística**

A UE tem 24 línguas oficiais, com dezenas de outras reconhecidas a nível nacional ou regional. No entanto, muitas destas línguas — mesmo algumas das 24 oficiais, como o dinamarquês, o finlandês, o sueco e o grego — representam menos de 0,6% do conteúdo disponível na internet. Outras, como o maltês, o irlandês, o estónio, o letão e o esloveno, são praticamente invisíveis online. Embora apenas 5% da população mundial fale inglês como primeira língua, o conteúdo em inglês representa metade da informação disponível na web, dominando os dados utilizados para treinar modelos de IA.



Imagem 1 — Conteúdo do Common Crawl por língua

Esta sub-representação digital tem consequências, uma vez que o treino dos LLMs dependem fortemente de conteúdo da web. Quando uma língua tem pouca presença online, corre o risco de ser excluída dos serviços de IA no futuro. Embora os modelos generalistas de maior dimensão consigam lidar com várias línguas, podem ainda assim falhar na captação da nuance linguística, do contexto cultural e da profundidade regional necessárias para aplicações verdadeiramente inclusivas. LLMs treinados com dados limitados são menos precisos, apresentam mais alucinações e erros, têm dificuldades com vocabulário e refletem mais enviesamentos.[[1]](#footnote-2)

Por exemplo, o Llama 3.1, um modelo *open source* bastante conhecido, apresenta uma diferença de desempenho superior a 15 pontos percentuais entre respostas em inglês e em grego, e uma diferença superior a 25 pontos entre inglês e letão. Isto significa que, se este modelo fosse um aluno do ensino secundário, estaria no topo da turma a inglês, a meio da tabela a grego e no fundo da turma a letão. E esta disparidade entre línguas é visível em todos os principais testes de desempenho de LLMs.[[2]](#footnote-3)



Imagem 2 — Desempenho no GSM8K vs CommonCrawl para línguas europeias com poucos recursos²

Em muitos casos, línguas com um património cultural profundo, como o bretão, o occitano e o romanche, que a UNESCO classifica como línguas em perigo, estão praticamente ausentes dos sistemas de IA mais utilizados atualmente.

**O poder económico da língua**

Este desenvolvimento desequilibrado dos modelos de linguagem tem consequências económicas reais. Quando os sistemas de IA não conseguem compreender ou responder na língua de uma região, limitam o acesso a serviços e oportunidades, prejudicando tanto as empresas locais como o crescimento económico em geral.

A difusão alargada da IA — a sua adoção e utilização em todas as economias — será um dos principais motores de inovação e crescimento da produtividade na próxima década. Tal como a eletricidade e outras tecnologias de uso generalizado no passado, a IA representa a próxima fase da industrialização.

Para as comunidades cujas línguas estão sub-representadas online, os benefícios da IA correm o risco de permanecer fora de alcance. Imaginem um empresário de um pequeno negócio em Malta que fala apenas maltês. Atualmente, as ferramentas avançadas de IA para tarefas como análise de mercado ou geração de conteúdo provavelmente não funcionam em maltês, limitando a forma como este empreendedor pode tirar partido da IA. Ou considerem um estudante polaco numa cidade fora de Varsóvia que não consegue encontrar recursos educativos baseados em IA na sua língua, o que pode afetar as suas oportunidades de aprendizagem. E mesmo quando uma plataforma de IA afirma suportar uma determinada língua, a experiência pode ser insatisfatória.

Os governos e instituições europeias reconheceram a importância de enfrentar esta situação. Para impulsionar a competitividade económica na era da IA, a Europa terá de derrubar as barreiras linguísticas e promover a difusão da IA em todo o continente. Segundo a Comissão Europeia, apenas 13,5% das empresas da UE utilizam IA. O Plano de Ação “AI Continent” da UE indica que eliminar as barreiras linguísticas no mercado único pode aumentar o comércio intraeuropeu em até 360 mil milhões de euros.

**Novas medidas para colmatar lacunas linguísticas**

Para ajudar a colmatar esta lacuna linguística, a Microsoft irá colaborar com parceiros europeus para aumentar a disponibilidade de dados multilingues. Em parceria com o Laboratório ICube da Universidade de Estrasburgo — uma instituição dedicada à engenharia, informática e imagem — iremos apoiar os esforços de treino de IA através da colocação de especialistas do Microsoft Open Innovation Center (MOIC) e do nosso AI for Good Lab em Estrasburgo, França. Esta equipa será apoiada por uma rede interna global de mais de 70 engenheiros, cientistas de dados e especialistas em políticas da Microsoft. Esta colaboração entre o MOIC, o AI for Good Lab da Microsoft e a Universidade de Estrasburgo irá também financiar dois investigadores pós-doutorados e disponibilizar até 1 milhão de dólares em créditos Azure.

Esta equipa começará por explorar o próprio repositório de dados multilingues da Microsoft, tornando-o acessível e transparente para o público europeu, incluindo os programadores *open source*. Isto inclui, por exemplo, dados de texto multilingue do GitHub e conjuntos de dados de voz. O MOIC e o GitHub irão colaborar com a [Hugging Face](https://huggingface.co/), uma plataforma popular de colaboração para o desenvolvimento de modelos de IA, para alojar e tornar os dados amplamente acessíveis. Esta iniciativa baseia-se na nossa relação existente com a [Hugging Face](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-deploy-models-from-huggingface?view=azureml-api-2) para disponibilizar uma série de modelos *open source* da sua coleção para implementação com um clique no catálogo de modelos do Azure. Isto inclui o lançamento, na semana passada, das mais recentes contribuições para a IA multilingue — o modelo [SmoILM3](https://huggingface.co/blog/smollm3), um modelo multilingue altamente eficiente com 3 mil milhões de parâmetros, que suporta 6 línguas: inglês, francês, espanhol, alemão, italiano e português.

O MOIC irá também colaborar com o [Common Crawl](https://commoncrawl.org/), um dos maiores repositórios gratuitos e abertos de dados recolhidos da web. O MOIC irá financiar trabalho no Common Crawl, recorrendo a nativos para anotar e introduzir dados em línguas europeias no conjunto de dados publicamente disponível do Common Crawl.

Além disso, o MOIC e o AI for Good Lab irão lançar um convite à apresentação de propostas para ajudar a expandir a oferta de conteúdo digital em 10 línguas europeias, disponibilizando as suas coleções de texto de forma responsável e ética, nos seus próprios termos, para o desenvolvimento e experiências de IA multilingue. As candidaturas a bolsas estarão disponíveis no [site do AI for Good Lab](https://www.microsoft.com/en-us/research/group/ai-for-good-research-lab/?msockid=2634a634d6416e992890b21bd7666fe9) a partir de 1 de setembro de 2025. Na seleção dos beneficiários, o MOIC e o AI for Good Lab irão focar-se em oportunidades para desbloquear dados em línguas com representação relativamente baixa no conteúdo online, como o estónio, o alsaciano, o eslovaco, o grego e o maltês. As bolsas incluirão créditos Azure e apoio técnico e de engenharia.

Embora mais dados multilingues sejam essenciais, melhores ferramentas tecnológicas e *know-how* também podem ajudar. Por exemplo, muitas línguas utilizam *scripts* (sistemas de escrita) que atualmente representam desafios para modelos originalmente concebidos para o alfabeto latino. Caracteres cirílicos, o alfabeto grego e a escrita cursiva do árabe têm todos propriedades diferentes. Os “tokenizers” genéricos frequentemente fragmentam estes sistemas de escrita de forma subótima, o que pode prejudicar a capacidade do modelo de aprender contexto a longo prazo ou ortografia correta nessas línguas. Novos avanços em técnicas que permitem a um modelo lidar com qualquer sistema de escrita de forma uniforme podem ser uma solução. Mecanismos mais eficazes para criar dados sintéticos e para processar e curar esses dados também podem ajudar, especialmente quando gerem eficazmente preocupações com privacidade e dados sensíveis.

O MOIC e o AI for Good Lab irão trabalhar para facilitar o desenvolvimento e a partilha de conhecimento, ferramentas e capacidades para enfrentar estes desafios e capacitar os programadores europeus. O AI for Good Lab irá publicar um guia detalhado sobre como criar conjuntos de dados linguísticos de alta qualidade e treinar LLMs locais para tirar o máximo partido dos dados existentes. Estes dois grupos irão também apoiar investigação relevante, organizar encontros, co-investir em projetos de dados comuns e garantir que o conhecimento, as ferramentas e as capacidades estão disponíveis onde são mais necessários. Estas equipas continuarão ainda a apoiar iniciativas como as do [Barcelona Supercomputing Center](https://nam06.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.bsc.es%2F&data=05%7C02%7Cgretchde%40microsoft.com%7C1977446b026d43989ade08ddbecda219%7C72f988bf86f141af91ab2d7cd011db47%7C1%7C0%7C638876512599706610%7CUnknown%7CTWFpbGZsb3d8eyJFbXB0eU1hcGkiOnRydWUsIlYiOiIwLjAuMDAwMCIsIlAiOiJXaW4zMiIsIkFOIjoiTWFpbCIsIldUIjoyfQ%3D%3D%7C0%7C%7C%7C&sdata=OFA8oVoKmVB94mblPVCAkCiqr%2FrriOcX0vNdR1z1pGo%3D&reserved=0), do [Basque Center for Language Technology](https://nam06.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.hitz.eus%2Fes%2Fhitz&data=05%7C02%7Cgretchde%40microsoft.com%7C1977446b026d43989ade08ddbecda219%7C72f988bf86f141af91ab2d7cd011db47%7C1%7C0%7C638876512599722987%7CUnknown%7CTWFpbGZsb3d8eyJFbXB0eU1hcGkiOnRydWUsIlYiOiIwLjAuMDAwMCIsIlAiOiJXaW4zMiIsIkFOIjoiTWFpbCIsIldUIjoyfQ%3D%3D%7C0%7C%7C%7C&sdata=qojBC9lOKoI9Ze3Dgx1z2gA%2FAMcrVfNmzlY7zus%2BCDI%3D&reserved=0) e da [Universidade de Santiago de Compostela](https://nam06.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnos.gal%2Fgl%2Fproxecto-nos&data=05%7C02%7Cgretchde%40microsoft.com%7C1977446b026d43989ade08ddbecda219%7C72f988bf86f141af91ab2d7cd011db47%7C1%7C0%7C638876512599733093%7CUnknown%7CTWFpbGZsb3d8eyJFbXB0eU1hcGkiOnRydWUsIlYiOiIwLjAuMDAwMCIsIlAiOiJXaW4zMiIsIkFOIjoiTWFpbCIsIldUIjoyfQ%3D%3D%7C0%7C%7C%7C&sdata=MZ8XT2CGWhOO6CLn%2BnG50e9q%2FPodhk0ahJk%2FXicgviA%3D&reserved=0), para lançar modelos de IA treinados em espanhol, catalão, basco e galego na Azure AI Foundry. Esta iniciativa permite aos programadores criar sistemas de IA que funcionem nas línguas oficiais de Espanha, promovendo a inovação e a inclusão.

Por fim, para promover a investigação responsável em IA e ajudar a reduzir a lacuna linguística, a Microsoft está a lançar duas novas colaborações académicas na Europa: uma com a Universidade de Estrasburgo e outra com a [IE University School of Science & Technology](https://www.ie.edu/school-science-technology/), em Espanha. O AI for Good Lab e o MOIC da Microsoft irão colaborar com a Universidade de Estrasburgo para conceder créditos Azure que apoiem investigação conjunta em IA. Na IE University, o AI for Good Lab irá também conceder créditos Azure para apoiar investigação centrada em línguas com poucos recursos, incluindo apoio a projetos finais de curso (capstone projects) que acelerem novas soluções focadas em linguagem e IA.

**Novas medidas para ajudar a proteger digitalmente o património cultural da Europa**

Desde 2019, a iniciativa Culture AI da Microsoft tem-se centrado na utilização da IA em todo o mundo para ajudar a preservar as línguas, os locais, as histórias e os artefactos que definem a história da humanidade. Impulsionada pelo AI for Good Lab e através de parcerias com organizações sem fins lucrativos, universidades, governos e instituições culturais, a iniciativa apoia projetos que digitalizam e protegem o património cultural — desde línguas em risco até marcos icónicos, incluindo em França, Roma e Grécia. Seja através da criação de réplicas digitais de locais históricos ou da disponibilização de coleções de museus, o objetivo é garantir que a identidade e diversidade culturais não só são preservadas, como também se tornam mais inclusivas e acessíveis na era digital.

Hoje anunciamos o nosso próximo projeto: a criação de uma réplica digital em parceria com o Ministério da Cultura francês e a empresa francesa [Iconem](https://iconem.com/). O projeto irá criar um gémeo digital de Notre-Dame, em Paris — um marco arquitetónico e cultural moldado ao longo de séculos. A construção de Notre-Dame começou em 1163 e prolongou-se por quase 200 anos, resultando numa obra-prima gótica com 128 metros de comprimento e duas torres com 69 metros de altura sobre o Sena. Após o devastador incêndio de 2019, Notre-Dame reabriu ao público no final de 2024. O projeto utilizará a tecnologia e os métodos que desenvolvemos com a Iconem para criar o [gémeo digital da Basílica de São Pedro](https://virtual.basilicasanpietro.va/en) no ano passado, com base em mais de 400.000 fotografias e algoritmos avançados de IA, em parceria com o Vaticano.

Tal como o projeto do ano passado documentou para o Vaticano cada detalhe da Basílica de São Pedro, este novo projeto irá criar uma réplica digital que preservará de forma permanente, em formato digital, todos os detalhes de Notre-Dame, garantindo que a sua estrutura, história e simbolismo são protegidos e acessíveis para as gerações futuras. Ao combinar imagem avançada com IA, iremos criar e doar ao Estado francês um gémeo digital que poderá ser utilizado por especialistas em conservação e exibido futuramente no Musée Notre-Dame de Paris.

Para além do projeto em Notre-Dame, anunciamos hoje também uma parceria com a [Bibliothèque Nationale de France](https://www.bnf.fr/fr), em colaboração com a Iconem, para digitalizar cerca de 1.500 maquetas cénicas de espetáculos da Ópera Nacional de Paris entre 1800 e 1914. As maquetas digitalizadas serão disponibilizadas através de experiências educativas e interativas, bem como sob a forma de um conjunto de dados acessível na plataforma Gallica da Bibliothèque Nationale de France, para projetos de investigação e IA cultural.

Por fim, estamos a iniciar um novo trabalho com o [Musée des Arts Décoratifs](https://madparis.fr/) para tornar publicamente acessíveis as descrições digitais detalhadas de aproximadamente 1,5 milhões de artefactos, desde a Idade Média até à atualidade. Este passo permitirá que investigadores em história, história da arte e conservação acedam a esta nova informação para estudo e utilização em investigação baseada em IA.

**Perspetivar o futuro: uma abordagem baseada em princípios**

Damos estes novos passos com humildade e respeito, reconhecendo que a preservação da diversidade linguística e cultural da Europa é uma tarefa dos europeus e para ser liderada por europeus. A União Europeia já lançou um esforço multilateral para reunir dados linguísticos da UE e digitalizar todos os tipos de património cultural. O nosso papel é contribuir e apoiar esses e outros esforços semelhantes. Nenhuma das iniciativas que hoje anunciamos irá gerar dados ou tecnologia proprietária para a Microsoft.

Em última análise, a melhor forma de capacitar mais pessoas em toda a Europa para responder a estas necessidades é dotá-las de competências em IA que lhes permitam ter sucesso nestas áreas. Tal [como concluiu recentemente a Comissão Europeia](https://digital-strategy.ec.europa.eu/en/news/eu-report-calls-member-states-accelerate-digitisation-cultural-heritage), a escassez de competências digitais no setor cultural está a dificultar os esforços de digitalização do património cultural em toda a Europa. Para ajudar a colmatar esta lacuna de competências, o MOIC e o AI for Good Lab irão partilhar o que sabem e aprendem sobre como realizar este trabalho essencial.

A tecnologia deve refletir a riqueza da humanidade — não apagá-la. Ao tomarmos medidas intencionais agora, podemos ajudar a garantir que a IA não apaga a diversidade linguística e cultural, mas sim a reforça. Este é um dos grandes desafios de equidade da era da IA. E se trabalharmos juntos — com propósito e urgência — podemos fechar esta lacuna e construir um futuro digital que honre todas as línguas, todas as culturas e todas as comunidades da Europa.

**CONTACTOS**

|  |  |  |
| --- | --- | --- |
| **Catarina Brito** | **Inês Filipe** |  |
| catarina.brito@lift.com.pt  | ines.filipe@lift.com.pt |  |
| 914 310 661 | 910 283 054 |  |

**SOBRE A MICROSOFT**

Microsoft (Nasdaq "MSFT" @microsoft) cria plataformas e ferramentas robustecidas por Inteligência Artificial para entregar soluções inovadoras que correspondam às necessidades crescentes dos nossos clientes. Enquanto empresa de tecnologia está comprometida com a democratização do acesso a Inteligência Artificial de forma responsável, mantendo a missão de capacitar todas as pessoas e organizações no planeta para atingir mais.

1. P. Rohera, C. Ginimav, G. Sawant e R. Joshi, “Será melhor perguntar em inglês? Avaliação da precisão factual de LLMs multilingues em inglês e em línguas com poucos recursos”, 28 de abril de 2025, arXiv: arXiv:2504.20022. doi: 10.48550/arXiv.2504.20022. [↑](#footnote-ref-2)
2. K. Thellmann et al., “Rumo à avaliação multilingue de LLMs para línguas europeias”, 17 de outubro de 2024, arXiv: arXiv:2410.08928. doi: 10.48550/arXiv.2410.08928. [↑](#footnote-ref-3)