Modelo de Linguagem em Grande Escala para a língua portuguesa

O Primeiro-Ministro anunciou o primeiro Modelo de Linguagem em Grande Escala de língua portuguesa de Portugal (LLM Português) no dia 11 de novembro de 2024.

Os LLM são modelos que utilizam Inteligência Artificial (IA) para processar, compreender e gerar texto em linguagem natural a partir de grandes quantidades de dados em diversos formatos (e.g. texto, imagem e vídeo). Estes modelos são componentes de vários tipos de sistemas, tais como sistemas de diálogo e *chatbots*, sistemas de pesquisa, sistemas automáticos de resposta a perguntas, entre outros.

Existem no mercado inúmeros LLM estrangeiros, que na sua grande maioria são modelos desenvolvidos por empresas privadas e otimizados para processar e gerar texto em língua inglesa, sendo que (i) apresentam um desempenho menos positivo no processamento e geração de texto noutras línguas e (ii) quando utilizados com dados sensíveis, reduzem a autonomia e soberania de dados, forçando que quem os utiliza tenha de partilhar os dados com estes fabricantes.

Vale a pena recordar que, nos últimos anos, têm sido frequentes as iniciativas em vários países de desenvolver LLM próprios, proficientes nas línguas dos países envolvidos, nomeadamente o ALIA, que fala castelhano, catalão, galego e basco, e do Viking 7B, que fala dinamarquês, finlandês, norueguês, islandês e sueco, entre outros exemplos.

Neste sentido, o Governo português anunciou ser uma prioridade o desenvolvimento e lançamento do primeiro LLM de língua portuguesa de Portugal, o AMÁLIA, Assistente Multimodal Automático de Linguagem com Inteligência Artificial. Esta iniciativa de desenvolvimento do LLM Português é a primeira divulgada no âmbito da Agenda Nacional de Inteligência Artificial que será apresentada de forma consolidada no 1.º trimestre de 2025.

O LLM Português AMÁLIA permitirá (i) contribuir para a preservação da soberania nacional, (ii) distinguir as diferentes variantes da língua portuguesa, (iii) reconhecer elementos da cultura e história de Portugal; (iv) permitir o controlo dos dados utilizados para a sua aprendizagem, e (v) assegurar condições de armazenamento e utilização de dados sensíveis, como é o caso da maioria dos dados da Administração Pública.

Esta é uma iniciativa do Governo português que será liderada conjuntamente pela Ministra da Juventude e Modernização, que tem competência delegada do Primeiro-Ministro relativamente à Inteligência Artificial, e pelo Ministro da

Educação, Ciência e Inovação.

Assim, a execução operacional desta iniciativa será liderada (i) pela Agência para a Modernização Administrativa (AMA, I.P.), que será responsável pela gestão da iniciativa e por assegurar as condições necessárias para a futura disseminação do LLM por todos os seus potenciais utilizadores públicos e privados, e (ii) pela Fundação para a Ciência e Tecnologia (FCT, I.P.), que será responsável por coordenar, junto dos centros de investigação, o treino e desenvolvimento do LLM, assegurar a infraestrutura necessária para o treino e alojamento do LLM, e pelo tratamento e curadoria dos dados que serão utilizados para este treino e desenvolvimento. Será com as infraestruturas e recursos humanos existentes nestas entidades que será possível executar uma iniciativa com objetivos e calendário ambiciosos.

O treino e desenvolvimento do AMÁLIA será executado por um consórcio liderado pelos centros de investigação Nova LINCS da Universidade Nova de Lisboa, Instituto de Telecomunicações e Instituto Superior Técnico, e integrará outros centros de investigação nacionais com reconhecido mérito no âmbito da Inteligência Artificial.

Assim, será possível aproveitar sinergias de projetos e investimentos já realizados, nomeadamente (i) os projetos de desenvolvimento do EuroLLM no Instituto de Telecomunicações e Instituto Superior Técnico, e do GlórIA e v-GlórIA no Nova LINCS, (ii) o projeto de curadoria dos dados do Arquivo.pt, que está a ser realizado pela FCT, I.P., e o (iii) o investimento realizado pelo Governo em infraestrutura de computação de alta-performance do Deucalion e Mare Nostrum 5.

Será também formado um Comité de Acompanhamento Especializado, constituído por peritos em Inteligência Artificial, como é disso exemplo o *Center for Responsible AI*. Este grupo será responsável por assegurar as melhores práticas de desenvolvimento de Modelos de Linguagem de Grande Escala, o cumprimento dos princípios éticos e de segurança e aconselhar sobre o potencial de aplicações do modelo nos diversos setores de atividade.

Esta iniciativa tem previsto um investimento de 5,5 milhões de euros e um calendário de trabalho e desenvolvimento de 18 meses, do qual resultará uma primeira versão multimodal do AMÁLIA. A este valor acresce o vasto investimento já realizado em infraestrutura de computação, projetos de desenvolvimento e recursos humanos especializados que contribuirão em grande medida para o desenvolvimento do LLM.

O financiamento necessário à concretização do LLM Português é assegurado no âmbito do Plano de Recuperação e Resiliência (PRR) e será desenvolvido

inteiramente por entidades públicas. O financiamento do projeto estará exclusivamente destinado às entidades públicas envolvidas no desenvolvimento do AMÁLIA.

Ao longo dos 18 meses, serão disponibilizadas diversas versões do AMÁLIA à medida que forem desenvolvidas novas funcionalidades:
No final do 1.º trimestre de 2025, será disponibilizado um AMÁLIA (versão beta); no final do 3.º trimestre de 2025 um AMÁLIA (versão base); e no final do 2.º trimestre de 2026 um AMÁLIA (versão multimodal).

Numa fase inicial, esta versão beta do AMÁLIA conseguirá receber e interpretar instruções em formato de texto e responder com base no conhecimento adquirido, também em texto escrito em português de Portugal.

Até ao final do 3.º trimestre de 2025, serão tratados novos dados sobre a língua, a cultura e história de Portugal. Estes dados serão provenientes de fontes como o Arquivo.pt, e serão utilizados para treinar o AMÁLIA na sua versão base. Só nesta versão será possível gerar respostas fiáveis e precisas sobre estas temáticas, bem como responder a questões com total segurança e sem risco para o utilizador. Nesta altura, o AMÁLIA já poderá ser integrado noutras aplicações externas e utilizar dados dessas fontes para gerar respostas de texto.

Todas as versões desenvolvidas serão disponibilizadas de forma gratuita e em *open source*, para que seja utilizado por todos, incluindo Academia, centros de investigação, entidades públicas, empresas e cidadãos. Para além das versões do LLM, todos os dados que suportam o treino serão disponibilizados em dados abertos, criando assim uma infraestrutura nacional de Inteligência Artificial que potencia o ecossistema de inovação da Inteligência Artificial em Portugal.

O LLM Português poderá ser aplicado a diversos domínios de atividade, sendo necessário afiná-lo e treiná-lo com dados específicos dos sectores de atuação, como Educação, Saúde, Serviços Públicos, entre outros.

No final dos 18 meses do primeiro projeto de desenvolvimento do LLM, o AMÁLIA versão multimodal já será capaz de interpretar diversos formatos de dados (e.g. texto, imagem e vídeo).

Esta versão final do LLM será diferenciadora na interpretação e geração de texto de língua portuguesa, no conhecimento que tem da literatura, cultura e história de Portugal. No entanto, o objetivo deste LLM não é de responder a perguntas genéricas em que o foco é a realização de raciocínios ou cálculos complexos, havendo outros LLM no mercado com bom desempenho nessas

tarefas.

Esta iniciativa vem reforçar a intenção do Governo português de colocar a tecnologia e o digital ao serviço das empresas, do Estado e das pessoas. O objetivo é posicionar o Estado como um acelerador e impulsionador do ecossistema de investigação e inovação no âmbito da Inteligência Artificial, e criar bases fundacionais importantes para o desenvolvimento da tecnologia através de processos criativos e transformadores em Portugal.

Assim, o AMÁLIA estará disponível para todos de forma aberta e gratuita, para que possam utilizá-lo para concretizar os seus projetos. Após este primeiro projeto, é lançado o repto a todos os utilizadores que partilhem as futuras evoluções do modelo e as coloquem ao serviço de todos os portugueses.