Contrato

“Criação do Modelo de Linguagem em Grande Escala da Língua Portuguesa de Portugal”

Inserido na medida RE-C05-i08 do Programa de Recuperação e Resiliência

Pela **Fundação para a Ciência e Tecnologia, I.P.**

Assinado de forma digital
por Maria Madalena Alves
Dados: 2025.03.10 19:52:02 Z

Pelo **Consórcio AMALIA,**

Assinado por: **JOSÉ JÚLIO ALVES ALFERES**
Num. de Identificação: [redacted]
Data: 2025.03.10 14:14:49 +0000
Certificado por: **Diário da República**
Atributos certificados: **Diretor da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa - Universidade NOVA de Lisboa**
CARTÃO DE CIDADÃO

Rogerio Colaço
Assinado de forma digital por Rogerio Colaço
Dados: 2025.03.10 15:02:41 Z

Assinado por: **ANA CRISTINA MOREIRA FREIRE**
Num. de Identificação: [redacted]
Data: 2025.03.10 15:32:07+00'00'
Certificado por: **Universidade do Porto**
Atributos certificados: **Diretor/a da Faculdade de Ciências da Universidade do Porto**
CARTÃO DE CIDADÃO

Assinado por: **Amílcar Celta Falcão Ramos Ferreira**
Num. de Identificação: [redacted]
Data: 2025.03.10 16:17:00+00'00'
Certificado por: **Diário da República**
Atributos certificados: **Reitor - Universidade de Coimbra**
CHAVE MÓVEL

Assinado por: **Sandra Cristina Almeida Paiva**
Num. de Identificação: [redacted]
Data: 2025.03.10 17:21:00+00'00'

PRR Plano de Recuperação e Resiliência
REPÚBLICA PORTUGUESA
Financiado pela União Europeia NextGenerationEU

Entre:

A **Fundação para a Ciência e a Tecnologia, I.P.**, adiante designada por FCT, I.P. com sede na Av. D. Carlos I, nº 126, 1249-074 Lisboa, com o número de identificação de pessoa coletiva 503 904 040, representada pela Professora Madalena Alves, na qualidade de Presidente do Conselho Diretivo da FCT, I.P., ao abrigo do artigo da 21.º, nº 3 da Lei n.º 3/2004, de 15 de Janeiro;

E

A **Universidade Nova de Lisboa**, parceiro do modelo *core,* coordenador financeiro e contratual, coordenador científico junto com o Instituto Superior Técnico e parceiro para o domínio específico da cultura e museus, através da sua Unidade Orgânica, Faculdade de Ciências e Tecnologia (NOVA FCT) com sede no Campus da Caparica, 2859-516 Caparica, com o número de identificação de pessoa coletiva 501559094, neste ato representada por José Júlio Alves Alferes, titular do Cartão do Cidadão nº 07377477, válido até 29-07-2029, na qualidade de Diretor, e

O **Instituto Superior Técnico**, parceiro do modelo *core*, coordenador científico junto com a Universidade Nova de Lisboa e parceiro para o domínio específico da fala, com sede na Av. Rovisco Pais, 1049-001 Lisboa, com o número de identificação de pessoa coletiva 501507930, neste ato representada por Rogério Colaço, titular do Cartão do Cidadão nº [redacted], válido até [redacted], na qualidade de Presidente, e

A **Universidade do Porto**, parceiro para o domínio específico de *media*, através da sua unidade orgânica Faculdade de Ciências da Universidade do Porto, com sede na Rua do Campo s/n, 4169-007 PORTO, com o número de identificação de pessoa coletiva 501413197, neste ato representada por Ana Cristina Moreira Freire, titular do Cartão do Cidadão nº [redacted], válido até [redacted], na qualidade de Diretora, e

A **Universidade de Coimbra**, parceiro para o domínio específico da ciência, com sede no Paço das Escolas, 3004-531 Coimbra, com o número de identificação de pessoa coletiva 501617582, neste ato representada por Amílcar Celta Falcão Ramos Ferreira, titular do Cartão do Cidadão nº [redacted], válido até [redacted], na qualidade de Reitor da Universidade de Coimbra, e

A **Universidade do Minho**, parceiro para o domínio específico da educação, com sede Largo do Paço, 4704-553 Braga, com o número de identificação de pessoa coletiva 502 011 378, neste ato representada por Sandra Cristina Almeida Paiva, titular do Cartão do Cidadão nº[redacted], válido até [redacted], na qualidade de Vice-Reitora, e

que integram o consórcio **AMALIA**, adiante designado por Consórcio.

É celebrado, ao abrigo do nº 5 do artigo 5ºA do Código dos Contratos Públicos, o presente contrato, verificadas as condições nele previstas para o estabelecimento de uma cooperação entre as partes.

Para o ano de 2025, a despesa inerente ao presente contrato está assegurada pelo compromisso n.º 920250000181 datado de 7 de março de 2025.

A despesa inerente ao presente contrato, para o ano de 2026 está inscrita no Sistema Central de Encargos Plurianuais (SCEP) da Direção-Geral do Orçamento (DGO), conforme declaração de compromisso indicada na PE/26/FCCN/2025, que aprovou a assunção de encargos plurianuais.

CLÁUSULA 1.ª

OBJECTO

O presente **CONTRATO** compreende os termos e condições para o estabelecimento da cooperação entre as partes tendo em vista a "**Criação do Modelo de Linguagem em Grande Escala da Língua Portuguesa de Portugal" – AMALIA** (*Automatic Multimodal Language Assistant with Artificial Intelligence*)".

CLÁUSULA 2.ª

OBRIGAÇÕES DAS PARTES

1. As partes obrigam-se a executar o **CONTRATO** em termos que se conformem com o nele estabelecido, nos anexos que dele fazem parte integrante e na legislação aplicável.
2. Para além de outras obrigações previstas na lei ou no presente **CONTRATO**, o as partes obrigam-se a:
   a) Assegurar que as atividades desenvolvidas cumprem as especificações técnicas acordadas;
   b) Cumprir os prazos estabelecidos para a execução das atividades previstas;
   c) Partilhar informação relevante para a boa condução dos trabalhos e alcance dos objetivos comuns;
   d) Garantir a confidencialidade das informações partilhadas no âmbito do contrato.

CLÁUSULA 3.ª

ESPECIFICAÇÕES TÉCNICAS

As partes obrigam-se a assegurar que a sua atividade no desenvolvimento da cooperação estabelecida pelo presente contrato obedece às especificações técnicas que constam do Anexo I ao presente **CONTRATO**, do qual fazem parte integrante.

CLÁUSULA 4.ª

PRAZOS

As partes obrigam-se ao pontual cumprimento de todos os prazos estabelecidos para a execução das prestações objeto do **CONTRATO**.

CLÁUSULA 5.ª

OBRIGAÇÃO DE PRESTAÇÃO DE INFORMAÇÃO

As partes obrigam-se a prestar, por escrito, toda a informação que lhes for solicitada relativa ao objeto da adjudicação ou à sua atuação em cumprimento das obrigações que para si decorrem do **CONTRATO**.

CLÁUSULA 6.ª

OBRIGAÇÃO DE SIGILO

A partes obrigam-se a não divulgar informações que obtenham em virtude da execução do **CONTRATO** durante a vigência deste e por um período de dois anos contados a partir da data da sua cessação.

CLÁUSULA 7.ª

FINANCIAMENTO E CONDIÇÕES DE ATRIBUIÇÃO

1. A FCT, I.P. afetará a quantia de 5.500.000,00 € (cinco milhões e quinhentos mil euros), acrescida de IVA à taxa legal em vigor, para desenvolvimento das atividades do consórcio, dos quais 900.000,00€ serão executados por si própria (Deucalion, Marenostrum, infraestrutura, desenvolvimentos de reforço do Arquivo.pt e pessoas), sendo a restante verba distribuída da seguinte forma:
   A) 2.475.000,00€ afetados à Universidade Nova de Lisboa (modelo *core*, pessoas, equipamento e domínio específico da cultura e museus)
   B) 1.000.000,00€ afetados ao Instituto Superior Técnico (modelo *core*, pessoas, equipamento e domínio específico da fala)
   C) 375.000€ afetados à Universidade do Porto (domínio específico de media)
   D) 375.000€ afetados à Universidade do Minho (domínio específico de educação)

E) 375.000€ afetados à Universidade de Coimbra (domínio específico de ciência)

2. As quantias previstas nas alíneas A) a E) do ponto anterior, serão satisfeitas através do pagamento de três faturas, emitidas pelo coordenador financeiro do consórcio, segundo o seguinte plano de faturação:

   a) A primeira fatura, correspondente a 40% da verba atribuída, é emitida após a entrada em vigor do contrato;
   b) A segunda fatura, correspondente a 35% da verba atribuída, é emitida após a entrega da primeira versão do modelo;
   c) A terceira fatura, correspondente a 25% da verba atribuída, é emitida após a aprovação da totalidade dos entregáveis acordados.

3. O pagamento referido no número anterior será efetuado ao coordenador financeiro do consórcio, que realizará a gestão e a distribuição da mesma aos restantes parceiros.
4. Poderão ser definidas condições de faturação diferentes das enumeradas no número anterior mediante acordo de ambas as partes.
5. As faturas a emitir pelo coordenador financeiro do consórcio assumem a forma de faturas eletrónicas, com os requisitos legais, nomeadamente os resultantes do artigo 299º-B do CCP.
6. As faturas referidas no número anterior serão pagas no prazo máximo de trinta dias a contar da sua receção.
7. A FCT I.P. utiliza a solução EDI e faturação eletrónica *ilink* (acessível em https://www.ilink.pt/), de registo gratuito, devendo todas as faturas emitidas pelo coordenador financeiro do consórcio [1].

CLÁUSULA 8.ª

VIGÊNCIA DO CONTRATO

1. O **CONTRATO** inicia a sua vigência na data da sua última assinatura e cessa vigência quando estiverem cumpridas todas as obrigações que dele resultam para os signatários ou a 30 de junho de 2026, consoante o que ocorrer primeiro.
2. A data de cessação de vigência do contrato pode ser posterior a 30 de junho de 2026, se o período de financiamento do PRR vier a ser alargado, caso em que será considerada a data-limite desse financiamento.

[1] Para qualquer questão de carregamento de faturas ou ligação/integração de sistema e de faturação deve ser contatada a iLink através do email apoio@ilink.pt ou pelo telefone 707 451 451.

3. A vigência da obrigação de confidencialidade estabelecida na cláusula 6.ª cessa na data em que cesse o prazo nele previsto.

CLÁUSULA 9.ª

RESOLUÇÃO PELAS PARTES

1. As partes podem resolver o **CONTRATO** com fundamento em incumprimento das obrigações previstas na cláusula 2.ª que determine a perda objetiva de interesse nas prestações que constituam o seu objeto.
2. A resolução do **CONTRATO** ao abrigo do disposto no número anterior determina a extinção dos créditos de que as partes sejam titulares, com exceção dos que se refiram a trabalho já realizado.

CLÁUSULA 10.ª

DESPESAS

1. Correm por conta das partes todas as despesas em que este haja de incorrer em virtude do cumprimento de obrigações emergentes do CONTRATO.

2. Corre por conta do consórcio o pagamento dos emolumentos, devidos ao Tribunal de Contas.

CLÁUSULA 11.ª

LEI APLICÁVEL

O **CONTRATO** rege-se pela lei portuguesa.

CLÁUSULA 12.ª

COMUNICAÇÕES

1. Para efeitos de comunicações relativas à fase de execução do **CONTRATO**, as partes podem recorrer aos seguintes meios de comunicação:
    a) Correio postal, através de carta registada ou de carta registada com aviso de receção;
    b) Correio eletrónico;

c) Outro meio de transmissão eletrónica de dados.

2. Todas as comunicações devem ser escritas e redigidas em língua portuguesa;
3. Para efeitos de estabelecimento das comunicações a que se refere a presente cláusula, as partes identificam os seguintes contactos, através dos quais as mesmas se devem concretizar:
   a) Pela FCT, I.P.:
      1. Nome do representante: [redacted]
      2. Endereço postal: Av. do Brasil, 101 1700-066 Lisboa
      3. Endereço eletrónico: [redacted]
   b) Pelo coordenador financeiro do consórcio:

      Execução das Atividades:
      1. Nome do representante: [redacted]
      2. Endereço postal: Quinta da Torre, 2829 -516 Caparica
      3. Endereço eletrónico: [redacted]

      Gestão:
      1. Nome do representante: [redacted]
      2. Endereço postal: Campus da Caparica, 2829-516 Caparica
      3. Endereço eletrónico: [redacted]

CLÁUSULA 13.ª

GESTORES DO CONTRATO

Para o exercício das funções de acompanhamento da execução do **CONTRATO** nos termos regulados pelo artigo 290º-A do Código dos Contratos Públicos são designados:

a) Pela FCT I.P.: [redacted]
b) Universidade Nova de Lisboa: [redacted]
c) Instituto Superior Técnico: [redacted]
d) Universidade do Porto: [redacted]
e) Universidade de Coimbra: [redacted]
f) Universidade do Minho: [redacted]

as partes.

CLÁUSULA 14ª

DISPONIBILIZAÇÃO DO MODELO DE LINGUAGEM EM GRANDE ESCALA DA LÍNGUA PORTUGUESA DE PORTUGAL - AMALIA

1. Todas as versões desenvolvidas do AMALIA ao longo da vigência do contrato serão disponibilizadas de forma gratuita e numa plataforma de LLM aberta (como por ex. *Hugging Face*), para que seja utilizado por todos, incluindo a academia, os centros de investigação, as entidades públicas, entidades privadas e os cidadãos.

2. O AMALIA poderá ser aplicado a diversos domínios de atividade, sendo necessário afiná-lo e treiná-lo com dados específicos dos setores de atuação (como a Educação, a Saúde ou os Serviços Públicos).

CLÁUSULA 15.ª

TRATAMENTO E PROTEÇÃO DE DADOS PESSOAIS

1. Para os fins estabelecidos nesta cláusula aplicam-se as disposições do Regulamento (UE) 2016/679 do Parlamento Europeu e do Conselho, de 27 de abril de 2016, relativo à proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados RGPD, bem como à legislação nacional e europeia aplicável em matéria de privacidade e proteção de dados.

2. Dando cumprimento ao disposto no número anterior, as Partes comprometem-se a avaliar as respetivas responsabilidades inerentes aos tratamentos de dados pessoais e, bem assim, celebrar os acordos necessários, nos quias definirão, nomeadamente, as finalidades de tratamento, as bases de licitude, os prazos de conservação, direitos dos titulares, procedimentos em caso de violação de dados pessoais, transferências internacionais de dados (se aplicável) e mecanismos de monitorização da conformidade com o RGPD e demais legislação relevante.

ANEXO I

ANEXO TÉCNICO

1. Introdução

O LLM Português contribuirá para dar às empresas nacionais, Administração Pública e demais entidades, o controlo sobre a língua portuguesa, permitindo que se exprimam corretamente nas várias variantes do português, capturando e preservando expressões idiomáticas, gírias e referências culturais próprias de Portugal.

Para além do valor cultural, importa considerar ainda a questão da soberania nacional em termos tecnológicos. Existem vários cenários de utilização de um LLM em que informação confidencial precisa de ficar em território nacional ou nas instalações do depositário dos dados. Ao contrário de LLMs comerciais, o AMALIA poderá ser utilizado em ambientes privados, evitando assim a transmissão dos dados para outro ambiente geograficamente diferente.

Propõe-se desenvolver o modelo em duas linhas de trabalho:

- Criação do modelo AMALIA base.
- Adaptações a Domínios Específicos e correspondente *benchmarking*.

Na primeira linha de trabalho, será criado o modelo base com capacidade para suportar linguagem e documentos visuais. Após a primeira versão do modelo AMALIA, terá início um conjunto de tarefas que visam especializar o AMALIA em domínios específicos.

A criação do modelo AMALIA será organizada em três grupos de atividades: (1) processamento de dados, (2) modelo de linguagem, (3) modelo multimodal.

2. Processamento de Dados

Para treinar o modelo AMALIA serão necessários dados em grande escala e de qualidade, na língua portuguesa. Serão utilizados métodos de recolha de dados e de filtragem de dados que têm como objetivo garantir que os dados processados são de qualidade, em português europeu e sem conteúdo perigoso ou não ético

3. Modelo de Linguagem

O treino do modelo de linguagem será realizado em larga escala e de forma distribuída. Em primeiro lugar o AMALIA será pré-treinado nos dados recolhidos. Posteriormente será treinado para ser alinhado com os valores humanos e para seguir instruções. Por fim, será feito um treino específico para garantir a segurança do modelo em termos de respostas geradas, mais concretamente, o modelo deverá rejeitar responder a pedidos sobre temas perigosos ou ilegais

3.1. Modelo Multimodal

Existem várias situações onde o AMALIA deverá ser capaz de responder a questões sobre conteúdo visual como documentos, imagens ou vídeo. Para garantir que o modelo consegue lidar com estes cenários, o modelo AMALIA será estendido com um componente de visão e seu adaptador para ligar a modalidade de texto e de imagem, ampliando as capacidades do modelo para além do processamento textual.

Será feito um treino de raiz do adaptador de modalidades e, posteriormente, dos modelos de ambas as modalidades. Após este treino inicial, iremos alinhar o modelo com humanos e torná-lo capaz de responder a perguntas sobre o conteúdo visual. Por fim, será igualmente feito um treino específico para garantir a segurança do modelo em termos de respostas geradas.

3.2. API para Integração

O modelo será disponibilizado de forma aberta em formato binário, em plataformas adequadas como por exemplo o *HuggingFace*, e será também implementada uma biblioteca de *software* para permitir acesso remoto ao LLM. A biblioteca de *software* suportará uma API por forma a permitir que cada entidade analise os seus requisitos específicos e opte pela forma mais adequada de operar e servir o modelo.

A biblioteca de *software* da API do AMALIA permitirá que se criem novas aplicações ou se integre em sistemas já existentes. Assim, tanto a AMA como outra entidade pública ou privada, terão a liberdade de decidir como se pretende operar do LLM AMALIA.

3.3. Entregáveis relativos ao Modelo *Core*

Mês	Grp. Ativ.	Entr.	Descrição
M3	GA1	E1.1	*Corpus* de dados para treino do modelo
	GA2	E2.1	LLM português
M6	GA4	E4.1	API para integração com fontes de dados
M9	GA3	E3.1	LLM português (público)
	GA1	E1.3	Enquadramento legal da utilização de dados para IA
M12	GA1	E1.2	*Corpus* de dados multimodais para treino do modelo
	GA2	E2.2	LLM português multimodal
	GA4	E4.2	API para integração com fontes de dados multimodais
M15	GA5	E5.1	Relatório de Validação
M18	GA3	E3.2	LLM português multimodal (público)

4. Plano de Atividades Detalhado

GA0: Acesso a recursos computacionais e dados

Parceiro: FCT I.P.

Para garantir a boa execução dos trabalhos, a FCT I.P. compromete-se a disponibilizar os seguintes recursos:

- Acesso a serviços de dados que permitam enriquecer o modelo com conhecimento sobre a variante da língua portuguesa europeia;
- Tempo de computação no Marensotrum 5;
- Tempo de computação no Deucalion;
- Aquisição, configuração, manutenção e operação de infraestrutura de *hardware* para suporte ao projecto na vertente do Arquivo.pt;
- Implantação, configuração e operação de infraestrutura de *software* para recolha e pré-processamento do *corpus* de dados para treino dos modelos, com base no Arquivo.pt;
- Apoio ao desenvolvimento da API para integração com fontes de dados;
- Extração, transformação e disponibilização de dados abertos para suporte a tarefas de I&D;
- Análise de fiabilidade e cibersegurança dos serviços desenvolvidos;
- Produção de documentação técnica e administrativa;
- Participação em ações de formação e disseminação junto da comunidade de ciência e tecnologia na área de Inteligência Artificial.

GA1: Preparação de Dados

Parceiros: Universidade Nova de Lisboa, Instituto Superior Técnico

Para treinar o modelo AMALIA serão necessários dados em grande escala e de qualidade, na língua portuguesa. Serão utilizados métodos de recolha de dados e de filtragem de dados que têm como objetivo garantir que os dados processados são de qualidade e não oferecem problemas de segurança.

No âmbito da tarefa A1.1 serão recolhidos dados de várias fontes, processados e organizados por tema e por qualidade. Estes dados serão utilizados na tarefa de treino causal e de instruções. Alguns dos dados utilizados em tarefas específicas noutras línguas serão traduzidas para português para treino do modelo.

No quadro da tarefa A1.2 será dedicada atenção à recolha e pré-processamento de *corpus* multimodais. Serão considerados *datasets* de referência traduzidos para português e serão recolhidos dados multimodais de outras fontes de informação. Neste contexto, serão considerados dados de imagem e vídeo.

A1.1 Recolha e pré-processamento de *corpus* de texto

- Integração dos dados do GLORIA com o EuroBlocks;
- Identificação de *corpus* com diversidade cultural relevante;

- Identificação de *corpus* de fontes relevantes;
- Tradução de *corpus* de referência.

A1.2 Recolha e pré-processamento de corpus multimodais

- Tradução de *corpus* multimodais;
- Identificação de *corpus* de fontes relevantes;
- Integração de dados multimodais.

A1.3 Análise legal da utilização de dados para IA

- Enquadramento legal da utilização de dados para IA
- Seleção da licença aberta do modelo

Entregáveis:

E1.1: *Corpus* de dados para treino do modelo;
E1.2: *Corpus* de dados multimodais para treino do modelo.
E1.3: Enquadramento legal da utilização de dados para IA

GA2: Treino do Modelo

Parceiros: Universidade Nova de Lisboa, Instituto Superior Técnico

O treino do modelo será feito por modelação causal de linguagem, por modelação de preferências e será ainda realizado treino em tarefas específicas para estender as capacidades de inferência do modelo. No quadro da tarefa A2.1 iremos utilizar treino causal onde o objetivo é aprender a prever o próximo *token* dado um contexto anterior. O treino é feito em modo auto-supervisionado, usando *corpora* massivos em português europeu (livros, artigos, páginas *web*). No âmbito da tarefa A.2.2 iremos utilizar técnicas de otimização direta de preferências (DPO), baseadas na aprendizagem de ordenação de respostas com base em preferências (humanas ou de outro modelo). Em resumo, a modelação causal de linguagem terá como objetivo ensinar o modelo a gerar texto de forma sequencial, enquanto a modelação de preferências ajustará o modelo para produzir respostas mais adequadas e alinhadas com os utilizadores.

Na tarefa A.2.3 serão consideradas várias tarefas de modelação de linguagem que permitem resolver várias operações de inferência em linguagem natural.

A2.1 Pré-treino

- Afinamento de hiperparâmetros;
- Treino e monitorização do modelo;
- *Annealing* e *super annealing;*
- Criação de *checkpoints* periódicos.

A2.2 Calibração por instruções e preferências

- Preparação de dados de instruções;
- Preparação de dados de preferências;
- Treino de modelo com DPO.

A2.3 Treino de tarefas *downstream core*

- Resposta automática a perguntas (QA), implicação textual (*entailment*), geração aumentada por recuperação (RAG), diálogo, tradução;
- Resposta Automática a Perguntas Visuais (VQA), descrição visual, pesquisa cross-modal.

Entregáveis:

E2.1: LLM português
E2.2: LLM português multimodal

GA3: Alinhamento e Confiança

Parceiros: Universidade Nova de Lisboa

Para mitigar os riscos e aumentar a utilidade prática do AMALIA, serão realizadas tarefas de alinhamento com princípios éticos para assegurar a confiança no modelo. Numa primeira atividade, a equipa irá aferir os vieses do modelo e o modelo será alinhado com segurança. Após este alinhamento, serão realizadas atividades de verificação da robustez do modelo através de *RedTeaming*, e, para mitigar as vulnerabilidades detetadas, o alinhamento com segurança será estendido.

A3.2 Alinhamento com segurança

- Identificação de vieses do modelo;
- Alinhamento por segurança.

A3.1 *RedTeaming*

- Implementação de ataques de *jailbreak* para testar robustez;
- Verificação e mitigação de falhas detetadas.

Entregáveis:

E3.1: LLM português (público)
E3.2: LLM português multimodal (público)

GA4: API e Experiência de Desenvolvimento

Parceiros: Universidade Nova de Lisboa

Para garantir uma utilização ágil e abrangente do modelo, serão criados vários artefactos que possibilitem o desenvolvimento rápido de aplicações e a sua utilização por terceiros.

Mais concretamente, a atividade A4.1 será focada na especificação da API para os casos de uso identificados como relevantes. Posteriormente, no quadro da atividade A4.2, será implementada uma biblioteca que disponibiliza a API para ser utilizada e operada por terceiros. Tendo em vista a rápida adoção do AMALIA, no contexto da atividade A4.3, a equipa irá criar um *website* com toda a documentação e tutoriais que ilustrem as capacidades do AMALIA. Por fim, na atividade A4.4, as atividades anteriores serão revistas e expandidas para o modelo multimodal.

A4.1 Design da API e experiência de desenvolvimento

- Comparação de soluções *cloud* para servir o modelo;
- Desenho da API para os casos de uso considerados.

A4.2 Desenvolvimento da API v1.0

- Implementação da API;
- Elaboração da documentação da API.

A4.3 Desenvolvimento do *website* da API

- Implementação de casos de uso da API;
- Criação de um *website* que disponilize o modelo, o *framework* da API e a correspondente documentação.

A4.4 Desenvolvimento da API v2.0 (multimodal)

- Desenho da API para casos de uso multimodais;
- Implementação da API para a versão multimodal do modelo;
- Implementação de casos de uso da API.

Entregáveis:

E4.1: API para integração com fontes de dados;
E4.2: API para integração com fontes de dados multimodais.

GA5: Validação e Verificação da Qualidade do Modelo

Parceiros: Universidade Nova de Lisboa, Instituto Superior Técnico

Tendo em vista uma validação prévia do modelo, iremos desenhar e implementar várias metodologias de avaliação da qualidade do modelo em vários domínios. Estas metodologias irão utilizar dados em português de alta qualidade, anotações de peritos e métricas que permitam aferir o progresso do modelo até à sua versão final.

A5.1 Avaliação do modelo geral

- Avaliação automática;
- Avaliação por linguistas.

A5.2 Tarefas de texto: QA, RAG, diálogo, tradução

- Avaliação automática;
- Avaliação por linguistas .

A5.3 Tarefas multimodais: VQA, descrição, pesquisa

- Avaliação automática;
- Avaliação por linguistas.

Entregáveis:

E5.1: Relatório de Validação

5. Capacidades e Domínios Específicos

5.1. GA6: Fala

Parceiros: Instituto Superior Técnico

O ambiente de LLMs AMALIA incluirá uma versão com capacidade para processar fala humana. Esta capacidade permitirá o desenvolvimento de aplicações que ofereçam interação de forma conversacional.

A6.1: Preparação de dados

Serão recolhidos dados de diversas fontes que contenham uma representatividade alargada de diferentes tipos de fala (leitura, conversação, etc.), falantes (idade, sexo, *status* social, etc.), domínios, condições acústicas e variedades dialetais do português europeu. Os dados serão normalizados e pré-processados de forma apropriada. Serão recolhidos e diferenciados dados sem anotações para efeitos de pré-treino e dados multimodais com anotações de texto para adaptação a tarefas finais, como reconhecimento automático de fala, tradução de fala para texto, sumarização de fala, classificação da intenção num turno de conversação ou deteção de discurso de ódio.

Fontes de dados: Repositórios de sessões de parlamentos (ARTV, EuroParl, etc.), tribunais, livros lidos, dados de rádio e TV, vídeo aulas e palestras (TEDX, EduCast), podcasts, etc.

A6.2: Codificador especializado para fala

Será treinado um modelo de tipo "*encoder*" utilizando objetivos auto-supervisionados apenas com os dados de fala, sem transcrições, podendo realizar-se o ajuste do codificador de fala a partir de modelos *open-source* multilingues pré-existentes. As capacidades multilingues do codificador são relevantes para tarefas finais multilingues, como por exemplo a tradução de fala para texto.

A6.3: Adaptador LLM para Codificador de Fala

Numa primeira fase, será treinado de raiz o adaptador de modalidade utilizando os dados anotados para reconhecimento automático de fala. O adaptador será treinado de forma supervisionada utilizando objetivos causais de predição do seguinte *token* para calibração na tarefa (instruction) de reconhecimento. Inicialmente, o codificador de fala e o modelo de linguagem *core* serão mantidos inalterados ou congelados. Numa segunda fase, o modelo de linguagem será também calibrado para incorporar fala.

A6.4: Relatório de validação

Será criado uma *benchmark* de teste para tarefas de fala em português europeu. A *benchmark* incluirá a tarefa de reconhecimento automático de fala e apresentará variedade de tipologia, domínio, condições acústicas, falantes e variedade dialetal. Adicionalmente, outras tarefas multimodais de fala e texto serão incluídas, como tradução de fala (estrangeira) para português (texto), sumarização de fala, classificação de intenções, deteção de discurso de ódio, tarefas pergunta-resposta sobre conteúdo multimodal, identificação de tópicos ou correção de pronúncia para estrangeiros. Esta *benchmark* será utilizada para avaliar a competência do modelo multimodal na tarefa de transcrição/reconhecimento de fala e nas outras tarefas adicionais.

A6.5: Modelo LLM-Encoder de Fala

A primeira versão do modelo multimodal será refinada mediante calibração de instruções, utilizando os dados anotados para as tarefas adicionais mencionadas. Nesta fase, apenas o decodificador do modelo de linguagem será calibrado.

Entregáveis:

Mês	Ativ.	Entr.	Descrição
M6	A6.1	E6.1	Recolha e processamento de dados de fala
M9	A6.1	E6.2	Recolha e processamento de dados de fala com anotações
M12	A6.3	E6.3	Modelo tipo *encoder* para portugês europeu
M15	A6.3	E6.4	Adaptador de modalidade para fala e calibração de instruções para reconhecimento
M15	A6.4	E6.5	*Benchmark* de fala
M18	A6.5	E6.6	Refinamento modelo multimodal com fala

5.2. GA7: Educação

Parceiros: Universidade do Minho

O AMALIA desempenhará um papel crucial na educação ao proporcionar às crianças e jovens a oportunidade de interagir com a IA num contexto que preserva e valoriza a identidade linguística e cultural portuguesa.

A7.1: Preparação de dados

O AMALIA será adaptado ao setor educativo, no domínio proposto, proporcionando suporte para alunos e professores. Esta versão especializada do AMALIA garantirá a preservação da identidade linguística e cultural portuguesa, promovendo um ambiente de aprendizagem enriquecido por inteligência artificial. Pretende-se um impacto significativo na forma como alunos e professores interagem com a tecnologia. O processo será composto pelas seguintes fases:

- Identificação e seleção de *corpus* educacionais relevantes (livros didáticos, artigos académicos, materiais pedagógicos);
- Adaptação de materiais de referência;
- Filtragem de dados para garantir qualidade, coerência e segurança;
- Integração eventual de fontes multimodais, como imagens explicativas, vídeos educativos e áudio;
- Estruturação dos dados para treino do modelo.

A7.2: Modelo especializado

Versão especializada do LLM AMALIA para educação, no domínio selecionado. Utilização de diferentes técnicas de aperfeiçoamento do modelo, nomeadamente diversos tipos de *fine-tuning, reinforcement learning with human feedback, few-shot learning, prompt engineering.*

Para otimizar a experiência de ensino, será adotado o conceito de *role-playing language models*, permitindo que o modelo atue com diferentes "personalidades". No contexto da educação, é fundamental que o LLM assuma o papel de um tutor que opere na zona de desenvolvimento proximal dos alunos e que forneça suporte adaptativo, incentivando a aprendizagem ativa. Além disso, esse tutor não apenas guia os alunos, mas também apresenta respostas diretas quando necessário. Embora a construção de um *dataset* específico para esse tipo de atuação seja um desafio, uma abordagem viável envolve o treino de modelos distintos, cada um especializado numa função diferente. A alternância entre esses modelos pode ser gerida através de um procedimento interno, permitindo a adaptação dinâmica às necessidades do aluno.

A7.3: Protótipo

Criação de protótipo de tutor virtual, com a capacidade de assumir uma ou várias "personalidades" distintas, de entre tutor interativo (zona de desenvolvimento proximal - perguntas abertas para guiar os alunos), tutor de resposta direta, tutor motivacional e tutor avaliador. O protótipo servirá como prova de conceito da utilização de um tutor num ambiente educativo, recorrendo ao modelo especializado da tarefa 2.

A7.4: Relatório de validação

Serão definidos e implementados protocolos de avaliação das funcionalidades, para aferir a eficácia do modelo especializado/protótipo. Pretende-se garantir os requisitos de qualidade, acessibilidade e impacto educacional esperados.

Entregáveis:

Mês	Ativ.	Entr.	Descrição
M9	A7.1	E7.1	*Corpus* educacional para o domínio selecionado (1ª versão).
M12	A7.2	E7.2	*Corpus* educacional estruturado para treino do modelo no domínio selecionado (versão consolidada)
M15	A7.3	E7.3	Versão especializada do LLM AMALIA para educação no domínio selecionado
M18	A7.3	E7.4	Prova de conceito – protótipo de tutor virtual
M18	A7.4	E7.5	Análise de eficácia e "*compliance*" do trabalho desenvolvido

5.3. GA8: Museus

Parceiros: Universidade Nova de Lisboa

O AMALIA irá permitir o acesso em português a informação sobre o património nacional, aumentando assim a disseminação e conhecimento sobre a cultura material dos museus.

A8.1: Preparação de dados

De forma a prover o AMALIA de conhecimento cultural português, nesta tarefa serão recolhidos e pré-processados dados referentes a aspetos culturais, bem como obras portuguesas. Serão consideradas duas tipologias de dados culturais: obras literárias representativas da literatura portuguesa, imagens e fotografias de bens culturais, expostos nos diversos museus portugueses, quer nacionais bem como regionais e especializados.

Serão considerados dados públicos da Biblioteca Nacional e dados multimodais de obras artísticas, em exposição em museus portugueses, mais concretamente os arquivos RAIZ (http://raiz.museusemonumentos.pt), e os dados recolhidos pelo PRR - Património Cultural 360 (https://arquiva.patrimoniocultural.gov.pt/patrimonio-cultural-360)..

A8.2: Modelo especializado

Nesta tarefa será efetuado o treino de um modelo especializado na descrição, anotação e contextualização de obras de arte em coleções museológicas, utilizando a coleção de dados provenientes da Tarefa 1. O modelo deverá suportar as seguintes funcionalidades:

- Contextualização histórica e artística, associando obras a estilos, períodos e influências;
- Identificação de passagens literárias;
- Geração de descrições semânticas a partir de imagens e atributos textuais;
- Anotação e identificação de elementos visuais relevantes.

Será utilizada uma abordagem de treino multitarefa, num regime de *fine-tuning*, de forma a obter um modelo especializado no agregado das funcionalidades. Será utilizado *instruction-tuning* para suportar a execução das funcionalidades mencionadas acima através de diálogo.

A8.3: Relatório de validação

Nesta tarefa serão definidos e implementados os protocolos de avaliação de cada uma das funcionalidades e vertentes, para avaliar a eficácia dos modelos desenvolvidos, assegurando a sua precisão, coerência e aplicabilidade no contexto cultural e museológico. Serão consideradas tarefas conversacionais tais como a) perguntas e respostas sobre cultura portuguesa e obras de arte, onde o objetivo será medir a capacidade do modelo compreender e contextualizar informação histórica e artística, b) identificação e citação de passagens relevantes de obras literárias, e c) anotação visual de obras, na qual se avalia a identificação de elementos presentes nas imagens, tais como estilos, temas e objetos culturais.

A8.4: Prova de conceito final

Nesta tarefa será libertada a versão final do modelo capaz de assimilar informação cultural, tanto da Biblioteca Nacional, como de museus nacionais. Será ainda criada uma prova de conceito que demonstre a capacidade do modelo AMALIA neste domínio.

Entregáveis:

Mês	Ativ.	Entr.	Descrição
M9	A8.1	E8.1	*Corpus* Multimodal Cultural Português
M12	A8.2	E8.2	Modelo especializado em tarefas de cultura
M15	A8.3	E8.3	Análise de eficácia dos modelos desenvolvidos
M18	A8.4	E8.4	Modelo final AMÁLIA com conhecimento literário e cultural

5.4. GA9: Ciência

Parceiros: Universidade de Coimbra

No domínio da Ciência, o AMALIA vai permitir otimizar a pesquisa de literatura em teses académicas, artigos de revista em português, documentos históricos e outros materiais. Esta aplicação específica do modelo beneficiará do repositório nacional de teses académicas e de repositórios locais das universidades portuguesas bem como de outros acervos em língua portuguesa de textos científicos.

A9.1: Recolha e Preparação de dados

Nesta tarefa serão identificadas fontes de literatura científica, considerando critérios como a quantidade de dados disponíveis, o seu formato, as áreas científicas abrangidas, e licenças de utilização. O ponto de partida será a exploração de repositórios digitais, como o Repositório Científico de Acesso Aberto de Portugal (https://www.rcaap.pt/) ou os repositórios científicos das instituições de ensino superior e investigação (e.g., estudo geral da Universidade de Coimbra https://estudogeral.uc.pt/), mas será também explorada a utilização de acervos de revistas científicas, com publicações em português, em diferentes áreas.

Com base nas fontes identificadas, será definido um conjunto de áreas científicas alvo. Documentos dessas áreas serão recolhidos e passarão por um processo de filtragem e curadoria, em que o texto será extraído e documentos duplicados e documentos escritos em línguas que não o português serão eliminados, . A última filtragem poderá ser feita com recurso a deteção automática de língua.

Os dados recolhidos serão integrados num corpo único que será utilizado para a especialização do modelo base no domínio da Ciência. O processo de recolha e compilação será documentado, o que incluirá uma análise da distribuição dos dados por fonte e áreas científicas.

A9.2: Modelo especializado

O modelo base será treinado tendo em vista a sua especialização no domínio da Ciência, utilizando a coleção de dados recolhida na Tarefa 1. O modelo deverá ser capaz de resumir literatura científica; de conseguir responder a perguntas em diferentes domínios científicos; e de ser capaz de explicar o raciocínio necessário para chegar às respostas geradas, através de *chain-of-thought* (CoT), sempre que possível, suportado por literatura existente.

Será seguida uma abordagem de *fine-tuning* continuado do modelo base, recorrendo a técnicas como *Direct Preference Optimization* (DPO) para a ordenação de respostas com base em preferências, ou *Iterative Reasoning Preference Optimization* (IRPO), mais adequado para CoT. Será realizado *instruction-tuning* para suportar a execução das funcionalidades mencionadas acima através de diálogo. Será ainda considerada a incorporação de *Retrieval Augmented Generation* (RAG) para suporte das respostas dadas.

A9.3: Validação

Nesta tarefa serão definidos e implementados os protocolos de avaliação do modelo especializado, de forma a quantificar a sua precisão, coerência e aplicabilidade ao contexto da Ciência. A avaliação terá como foco as tarefas de: (a) resposta automática a perguntas em diferentes domínios científicos; (b) explicação do raciocínio, onde poderão ser citadas passagens relevantes de literatura científica; (c) sumarização automática de um ou mais documentos.

De forma a identificar dados de referência a usar na avaliação, será feita uma análise de conjuntos de dados públicos, já compilados, e definidas adaptações necessárias à sua

utilização. Será também considerada a compilação de novos dados de avaliação recorrendo, por exemplo, a exames de instituições de ensino em português ou a fóruns científicos na *web*.

A9.4: Prova de conceito

Nesta tarefa será disponibilizada a versão final do modelo. A tarefa envolverá a criação de uma prova de conceito que demonstre a capacidade do modelo especializado no domínio da Ciência, cujo desempenho será comparado com o do modelo base.

Entregáveis:

Mês	Ativ.	Entr.	Descrição
M3	A9.1	E9.1	Coleção de dados no domínio da Ciência
M12	A9.2	E9.2	Primeira versão do modelo de linguagem especializado no domínio da Ciência
M15	A10.15	E10.3	Relatório de validação
M18	A10.18	E10.4	Modelo final especializado no domínio da Ciência

5.5. GA10: Média

Parceiros: Universidade do Porto

A10.1: Recolha e Preparação de dados

Nesta tarefa, serão recolhidos e pré-processados dados no domínio dos *media*, obtidos a partir de diversas fontes, com o objetivo de especializar o modelo em tarefas como a sumarização abstrativa, a identificação e justificação de narrativas dominantes em artigos noticiosos de natureza manipuladora, a deteção de técnicas de persuasão em conteúdos jornalísticos, a geração automática de notícias com base na deteção de tópicos em atas municipais e a criação de artigos desportivos a partir de estatísticas.

Para a recolha de dados, prevê-se a utilização de *corpora* previamente anotados, públicos e resultantes de investigações conduzidas pelo grupo de investigação, nomeadamente na extração e caracterização de narrativas e técnicas de persuasão. Além disso, serão estabelecidos ou estendidos protocolos com entidades parceiras, como por exemplo a Agência Lusa, o ZeroZero.pt e diversas Câmaras Municipais. Estes protocolos contribuirão para a expansão do conjunto de dados disponível, que passarão por processos de anotação, curadoria e tradução.

A10.2: Modelo para geração de notícias

Ajuste fino (*fine tuning*) de um modelo especializado na geração de notícias, utilizando como base os *corpora* identificados e coletados na Tarefa 1. O modelo deverá suportar a geração de notícias a partir de diferentes tipos de entrada, incluindo estatísticas

desportivas, texto não jornalístico e atas municipais, bem como a criação de infografias baseadas em discurso jornalístico e a geração de legendas para fotografias. O ajuste fino será realizado a partir do modelo core AMALIA ou de modelos *open-source* multilingues alternativos (e já existentes), tendo em conta a necessidade de assegurar a coerência narrativa dos textos gerados no domínio do jornalismo e da variedade português europeu.

A10.3: Modelo para análise de notícias

Ajuste fino de um modelo especializado na compreensão de notícias, com foco na deteção de discurso manipulador, identificação e justificação de narrativas dominantes, extração de elementos narrativos, deteção de posicionamento político e identificação de técnicas de persuasão em textos jornalísticos, extração de eventos em notícias desportivas. O modelo resultante será capaz de representar semanticamente os textos jornalísticos, capturando padrões discursivos, estrutura narrativa e sinais de viés ou persuasão. O modelo será treinado com base em *corpora* anotados previamente na Tarefa 1, incluindo dados de fontes jornalísticas diversas e conteúdos anotados quanto à natureza persuasiva e manipuladora. A adaptação do modelo envolve a deteção de discurso manipulador, a identificação e justificação de narrativas dominantes, a extração de elementos narrativos, a deteção de posicionamento político e a identificação de técnicas de persuasão. O ajuste fino será realizado a partir do modelo core AMALIA ou de modelos *open-source* adequados para as tarefas em questão.

A10.4: Relatório de validação no segmento media

Desenvolvimento de um *benchmark* de avaliação para validar o desempenho dos modelos. O *benchmark* incluirá a definição de modelos base de referência baseadas no estado da arte e a seleção de métricas de avaliação adequadas para cada tarefa. Serão conduzidas avaliações específicas para medir a capacidade dos modelos nas seguintes tarefas: (1) identificação de tópicos relevantes, garantindo que os modelos identificam informação essencial a partir de um texto dado como *input*; (2) geração de textos jornalísticos na variedade português europeu a partir de fontes diversas; (3) extração e análise de elementos narrativos, como personagens e eventos; deteção de técnicas de persuasão e discurso manipulador; (4) justificação de narrativas dominantes, verificando a capacidade dos modelos em explicar tendências narrativas.

O relatório final incluirá uma análise detalhada dos resultados, identificando pontos fortes e limitações dos modelos.

A10.5: Prova de conceito

Será desenvolvida uma prova de conceito que demonstre a capacidade do modelo AMALIA neste domínio, como por exemplo um RAG para a geração de notícias. Possíveis melhorias obtidas no modelo especializado serão integradas no modelo *core*. Eventuais limitações identificadas na tarefa 4 serão mitigadas.

Entregáveis:

Mês	Ativ.	Entr.	Descrição
M6	A10.1	E10.1	Conjuntos de dados anotados e armazenados em suporte adequado para o processamento de treino e avaliação e a sua caracterização.
M9	A10.2	E10.2	Modelo especializado para geração de notícias em português europeu
M12	A10.3	E10.3	Modelo especializado para compreensão de notícias e análise de narrativas
M15	A10.4	E10.4	Relatório de validação do desempenho dos modelos em várias dimensões
M18	A10.5	E10.5	Prova de conceito