Projeto IFIC

Descrição Técnica

Ricardo Pinto

rpinto@moongy.pt

Índice

1 Resumo Executivo

1.1 Oportunidade de Mercado

A inteligência artificial generativa está a tornar-se no elemento diferenciador de competitividade, a acumular investimentos de centenas de milhares de milhões de euros e com crescimento continuado previsto para os próximos anos. Alvo de imenso investimento e investigação, a mesma tem sido rapidamente melhorada e começa já a reformular cadeias de valor em vários setores. A maturação destes modelos de IA generativa, tal como a explosão de serviços que a fornecem e de modelos open-source que a disponibilizam, obrigam a um planeamento estratégico rápido das nações e das suas empresas, resultando numa janela de oportunidade muito interessante para a construção de novos produtos e serviços alavancados sobre esta tecnologia.

Um dos potenciais latentes destes modelos de última geração, é o de gerar conteúdos de alta-fidelidade nos formatos de texto, áudio e vídeo, permitindo uma possível construção de "experiências digitais personalizadas e interativas", as quais cremos virem a ser altamente comercializadas num futuro próximo. A procura por este tipo de experiências existe de facto, a adoção imediata do ChatGPT é prova disso, e irá crescer rapidamente nos setores onde as mesmas têm enquadramento direto, como em vendas, marketing, ensino, entretenimento, cultura, serviços públicos, entre muitos outros. As organizações irão necessitar de ferramentas de produção de conteúdos desta natureza, de forma fácil, organizada e fiável, e que garantam transparência sobre o processo criativo e a sua originalidade intelectual.

Neste contexto, existe uma oportunidade clara de mercado para plataformas que forneçam ferramentas que permitam uma fácil construção de experiências personalizáveis e interativas, para a sua respetiva divulgação e comercialização, e que garantam elevados padrões de qualidade e de fidelidade.

1.2 Proposta de Projeto

Pretende-se investigar e desenvolver uma plataforma para a criação de experiências digitais personalizáveis e interativas, multimodais de natureza, capaz de criar texto, áudio e vídeo, tanto com elementos produzidos em tempo real, como com elementos produzidos antecipadamente, os quais seguem contextos, histórias e narrativas definidas pelos editores dos conteúdos, mas que incorporem flexibilidade para se adaptarem dinamicamente aos consumidores dos conteúdos.

Esta plataforma permitirá compor experiências para diversas áreas distintas, como educação e formação, marketing e publicidade, entretenimento e cultura, saúde e cidadania, entre muitas outras. Embora aplicável a diversos setores, o âmbito do projeto incidirá nas áreas de vendas e de apoio ao consumidor, não obstante garantindo a flexibilidade para outros verticais funcionais serem acrescentados. A nível de planeamento de desenvolvimentos do projeto, pretendemos ter duas macro fases:

Desenvolvimento de Piloto

Na primeira fase, pretendemos desenvolver um piloto que demonstre o potencial da plataforma, para imediatamente confirmarmos a sua adequação ao mercado e procedermos a ajustes e melhoramentos.

Este deve refletir as características fundamentais da plataforma, as quais constituem as linhas de investigação principais deste projeto, nomeadamente:

- Geração controlada e ancorada de experiências
- Diálogo empático com análise de emoções
- Construção de narrativas interativas

Incorporado destas características base, o piloto deverá permitir a composição de interações de prospeção em texto e voz, seguindo uma narrativa pré-definida, com a capacidade de ajustar-se às interações do cliente, permitindo posterior seguimento por um comercial humano, mediante integração com um sistema de registo de prospeção qualificada.

Desenvolvimento de Produto Mínimo Viável

Na segunda fase, pretendemos desenvolver um produto cloud escalável em regime SaaS, que forneça as funcionalidades principais de geração de experiências, pronto para ser comercializado e integrável com sistemas de clientes para ativar processos customizados pelos mesmos. A título de exemplo, uma interação com o comercial inteligente de uma operadora de telecomunicações, que elucida, clarifica e apresenta os produtos e serviços mais adequados ao cliente, e que fecha a venda, pode depois automaticamente integrar com os sistemas internos e proceder ao envio da documentação para contratualização do serviço e agendamento de instalação do equipamento contratado.

O resultado esperado dos desenvolvimentos será uma plataforma funcionalmente coerente e extensível, com as devidas ferramentas para criação, edição e publicação de experiências imersivas, que habilitem profissionais a expandir o seu processo criativo para novos modelos de comunicação e interação.

1.3 Adequação do Projeto ao Mercado

A utilização de IA generativa para providenciar serviços de valor acrescentado é uma área em franca expansão com imenso potencial por explorar. A ideia é clara: oferecer serviços de inteligência especializada para um dado âmbito específico. O desafio encontra-se no ajuste da ideia ao mercado e na execução bem-sucedida da ideia. Do levantamento de mercado que fizemos, existem algumas empresas de destaque pioneiras na área de atuação de vendas ou de apoio ao consumidor, ainda que em número reduzido e de dimensão pequena.

Todas as empresas que identificámos (ver Tabela 1) são startups e PMEs, curiosamente duas em Portugal, a Agentifai [1] e a Visor [2]. Ambas focam-se em fornecer serviços de apoio ao cliente e já têm alguns clientes de dimensão interessante a nível nacional. A nível internacional destacam-se a Pitch Avatar [3] para criar apresentações interativas, a D-ID [4] para criar apresentações interativas com avatares digitais, a Uneeq [5] para o treino de vendas através da interação com avatares

virtuais, a Synthesia [6] para a criação de avatares para formação e marketing, e a HeyGen [7] também para a criação de avatares interativos.

Da nossa análise, a faturação anual conjunta destas empresas ronda apenas os 300M USD, grande parte desta concentrada na Synthesia e na HeyGen, claramente demonstrando que existe uma procura forte por este tipo de serviços e que o mercado está muito distante de estar saturado, considerando o investimento colossal que está a ser feito na área.

Analisando cada um dos serviços em maior detalhe, podemos notar que estas plataformas ainda são bastante rudimentares e com funcionalidades simples. A título de exemplo, tanto a Synthesia como a HeyGen oferecem serviços simples e exclusivos à criação e customização de avatares. O seu foco é exclusivamente em apresentar um avatar realista e dotá-lo de discurso. O resultado final é bastante impressionante, o que mais peso dá à tecnologia, mas muito distante do verdadeiro potencial desta área tecnológica – a criação de experiências multimodais, emotivas, personalizadas e interativas.

Efetivamente, nenhuma das plataformas oferece funcionalidades de customização avançada, e nenhuma incorpora com propriedade as características fundamentais que constituem as nossas linhas de investigação.

Adicionalmente, identificámos também as seguintes lacunas e oportunidades de mercado:

- Integração e sequenciação de modos: nenhuma das plataformas explora a criação de experiências que permitam saltar entre diferentes modos, como texto, para áudio, para audiovisual, pré-geradas ou em tempo real. Histórias podem ser apresentadas em diferentes formatos distintos, apelando também a diferentes tipos de atenção e interação.
- Deteção de emoções: a maioria das soluções não referem deteção de emoções, e as que referem não oferecem nenhuma descrição aprofundada sobre o serviço, indicando que se trata também de um complemento muito rudimentar.
- Equilíbrio entre modalidades e custo: nem todas as plataformas apresentam simuladores para cálculo de custo, e como são muito específicas a um certo formato, nenhuma permite que utilizador escolha diferentes formatos de comunicação para otimizar a relação custo-benefício das suas experiências.
- Personalização e facilidade de uso: nenhuma das plataformas apresenta funcionalidades avançadas de personalização de experiências. São todas muito focadas no seu caso de uso, e muito simplistas ainda.
- Especialização em vendas: nenhuma das plataformas é especializada em vendas. Têm a capacidade de criar processos rudimentares de fluxos de interação que podem teoricamente ser aplicados a vendas, mas que necessitam de ser muito melhorados para ter algum tipo de sucesso. Estamos a referir-nos a comerciais virtuais, que são capazes de fazer recomendações personalizadas, guiando o utilizador à escolha do produto adequado para si, de serem capazes de negociar, e de conseguir fazer o fecho do negócio, através da integração com sistemas terceiros para pagamento e encomenda de produto ou para envio de contratualizações de serviço.
- Integração com ferramentas: a integração destas plataformas com serviços é pouco sofisticada. Alguns são pouco claros a nível de integrações disponíveis, como a Agentifai, a

Visor e a Pitch Avatar; a D-ID oferece uma lista de integrações mas mais orientadas à publicação dos seus conteúdos do que à integração com sistemas; a Uneeq oferece um SDK para permitir a configuração pelo cliente de integrações customizadas; e a Syntesia oferece uma lista de integrações com plataformas externas. A HeyGen segue uma abordagem diferente, e muito interessante, na medida em que permite a integração com a plataforma n8n – um orquestrador de execução cloud, delegando a complexidade da configuração de fluxos de execução para o mesmo.

- Segurança da IA: nenhuma das ferramentas apresenta soluções para validar se os seus sistemas inteligentes não estão a ter halucinações, para garantir que são factuais e eticamente corretos. Há uma falta clara de análise da qualidade destes sistemas, algo que é fundamental para empresas com maior dimensão ou destaque.

Os pontos supra identificados são demonstrativos de que a exploração deste mercado ainda está muito nascente, com inúmeras oportunidades de melhoria e de crescimento. A nossa ideia de projeto, de estratégia de desenvolvimento e de ia ao mercado, coloca-nos numa posição interessante relativamente à concorrência, no sentido em que pretendemos, por um lado, desenvolver uma ferramenta generalista para a criação de experiências multimodais, emotivas, personalizadas e interativas, mas, por outro lado, com foco em vendas e apoio ao consumidor, e com forte integração com serviços externos para a execução de processos.

Empresa	Descrição do Produto
Agentifai	Assistente virtual voice-first para customer care, combinando IA conversacional e IA generativa para automatizar atendimentos. Foca em UX natural por voz e texto.
Visor.ai	Plataforma de automação de atendimento ao cliente (no-code), com assistentes virtuais em chat e voz alimentados por IA. Foco em melhorar rapidez e qualidade no suporte via IA.
Pitch Avatar	Plataforma para criar apresentações interativas com avatar de IA. Gera automaticamente guião, voz e apresentador virtual que conduz apresentações 24/7 de forma personalizada. Permite Q&A em tempo real com espectadores.
D-ID	Plataforma de avatares digitais realistas para interação face-a-face. Permite criar assistentes virtuais que olham, falam e expressam emoções como humanos, incorporando a personalidade da marca.
UneeQ	Plataforma empresarial para criar humanos digitais – avatares animados que conversam com utilizadores de forma natural. Projetada para escalabilidade, oferece experiências interativas realistas em marketing, vendas, suporte e treinamento.
Synthesia	Plataforma líder em geração de vídeos por IA. Permite criar vídeos com apresentadores virtuais apenas fornecendo um texto – sem necessidade de filmagens. Oferece avatares-falantes personalizáveis (inclusive clones digitais de pessoas reais) e múltiplas vozes/línguas.
HeyGen	Plataforma de geração de vídeos por IA orientada ao conteúdo comercial. Permite criar avatares falantes a partir de fotos ou texto em poucos cliques, com vozes realistas e movimento labial sincronizado. Oferece biblioteca de avatares prontos e opção de clonar a própria imagem.

Tabela 1 - Listagem das empresas analisadas e das suas respetivas áreas de atuação principais

1.4 Capacidade de Implementação

Resumo sobre a capacidade da empresa em implementar o projeto

- Histórico da empresa
- Demonstração de conhecimentos
- Potenciais clientes

1.5 Retorno sobre o Investimento

O retorno sobre o investimento (ROI) para este projeto é projetado como altamente atrativo, considerando o crescimento exponencial do mercado de IA generativa e conversacional. Com base em dados de mercado, o setor de IA generativa foi avaliado em cerca de 16-21 mil milhões USD em 2025 e é projetado para atingir 97-109 mil milhões USD até 2030, com um CAGR de 35-37%. Para a IA conversacional, o mercado é estimado em 11-14 mil milhões USD em 2024-2025, crescendo para 41-50 mil milhões USD em 2030, com CAGR de 23-35%. Assumindo um investimento inicial de 630.000 EUR (9.000 horas a 70 EUR/hora), com 60% financiado pelo PRR (investimento líquido de 252.000 EUR), o ROI em 4 anos pode exceder 100%, impulsionado por receitas SaaS escaláveis e baixa marginalidade de custos após o desenvolvimento.

1.5.1 · Planeamento de Ida ao Mercado

O plano de go-to-market (GTM) segue estratégias comprovadas para plataformas AI SaaS, como as adotadas pela Synthesia e HeyGen, que priorizam freemium para aquisição rápida e upselling para retenção. Dividido em fases:

- **Fase 1: Lançamento do Piloto (Q1-Q2 2026):** Foco em MVP para verticais de vendas e apoio ao consumidor, usando content marketing (webinars, demos interativos) e parcerias com empresas de telecomunicações e retalho para pilots gratuitos. Estratégia de inbound via LinkedIn e eventos como Web Summit para atrair early adopters.
- **Fase 2: Escala SaaS (Q3 2026-Q4 2027):** Transição para modelo de subscrição (freemium básico + premium a 50-200 EUR/mês por utilizador), com integrações API para CRM (ex.: Salesforce). Campanhas paid ads no Google Ads e LinkedIn, visando PMEs europeias. Parcerias com integradores (ex.: consultoras como Deloitte) para expansão B2B.
- **Fase 3: Expansão Global (2028-2030):** Localização multilingue e entrada em mercados EU/USA via alianças (ex.: AWS Marketplace). Métricas chave: CAC < 200 EUR, LTV > 2.000 EUR/ano, churn <10%. Esta abordagem, inspirada em GTM AI-driven (ex.: HockeyStack para automação workflows), garante aquisição eficiente e retenção via atualizações contínuas.

1.5.2 · Projeção de Vendas

Projeções conservadoras assumem captura de 0,1-0,5% do submercado conversacional AI (foco em verticais de vendas/apoio), com ARPU de 1.000 EUR/ano (subscrições SaaS) e crescimento alinhado ao CAGR de 23-35%:

- **Ano 1 (2026):** 50.000 EUR (50 clientes iniciais via pilots, foco EU; churn 15%).
- **Ano 2 (2027):** 100.000 EUR (150 clientes, expansão via parcerias; churn 10%).
- **Ano 3 (2028):** 200.000 EUR (300 clientes, entrada USA; churn 8%).
- **Ano 4 (2029):** 300.000 EUR (500 clientes, escala global; churn 5%). Cumulativo até 2030: ~3 mil milhões USD no mercado global; nossa quota conservadora gera ROI >1.000% pós-investimento líquido. Break-even no Ano 4, payback em 3,8 anos, impulsionado por margens >70% em SaaS maduro

2 Introdução

2.1 Contexto Tecnológico

Nas últimas décadas, a inteligência artificial evoluiu de sistemas arcaicos baseados em regras fixas, para sistemas estatísticos do domínio da ciência de dados, até aos atuais sistemas de arquiteturas de redes neuronais profundas, que permitem modelar e entender padrões complexos [8].

Esta última evolução, catalisada pela descoberta das arquiteturas transformer [9], popularizadas com o advento do ChatGPT [10], [11], redefiniu o estado de arte na produção e processamento de conteúdos. Recentemente, estes modelos foram generalizados para produzir diferentes formatos, permitindo a geração de texto, áudio, imagens e vídeo de uma forma simples e unificada [12]. Esta generalização consistiu num passo importante para a transformação de LLMs (Large Language Models) para LMMs (Large Multimodal Models), assumindo assim uma unificação de capacidades de percepção e geração [13].

Paralelamente, surge uma explosão de modelos abertos, muitos publicados na plataforma Hugging Face [14] e executáveis transparentemente através de ferramentas como o Ollama [15], desde que o hardware local o permita. O ecossistema open-source, promove também iniciativas para o treino e ajuste fino de modelos, transformando esta disciplina de uma fechada e pouco conhecida para o centro da nova indústria criativa do futuro.

A IA generativa permite agora uma criação personalizada de conteúdos a uma velocidade, preço e qualidade sem precedentes, representando um enorme salto em produtividade e uma mudança de paradigma em como o conhecimento é produzido e transmitido. Este salto constitui um avanço na economia digital com um impacto profundo, potenciando a aceleração na criação de conteúdo e na automação de processos cognitivos, resultando num aumento transversal de produtividade em domínios como o design, publicidade, produção audiovisual, formação, engenharia, educação, entre muitos outros. Tarefas que recentemente eram morosas e necessitavam da participação de peritos, podem agora ser executadas em minutos, permitindo uma rápida prototipagem, experimentação e iteração, e eventual produção de resultados, com uma velocidade e precisão sem precedentes.

Consequentemente, a estrutura do tecido laboral especializado está a adaptar-se rapidamente. Novas profissões e especializações estão a surgir, como engenheiro de prompts, curador de IA, designers de media sintético, entre outras profissões completamente fora do nosso quotidiano. Estas profissões representam uma mudança de paradigma na forma de trabalhar, indicando que o futuro estará intimamente interligado com sistemas generativos, onde os humanos deixam de produzir diretamente para, em vez disso, controlarem e direcionarem a produção automática, provinda destes novos sistemas. Uma forma mais avançada deste conceito surge também com a introdução de agentes cognitivos, sistemas que colaboram em atividades de direção, planeamento e produção. Tornam-se colaboradores que estendem e complementam equipas humanas [16], [17].

Alinhada com a evolução rápida desta área, surge a proposta que será detalhada neste documento, a qual consiste na utilização destas novas tecnologias para a criação de experiências inteligentes e interativas, assentes sobre conteúdos fiáveis e ancorados em informação fidedigna. Esta iniciativa de investigação propõe desenvolver aplicações e ferramentas que permitam a construção fácil de experiências multimodais, tanto expositivas como interativas e com aplicação em diversas áreas, como vendas, marketing, formação, publicidade, entre outras, e que utilizem LMMs como base generativa para as mesmas.

Todas estas novas tecnologias trazem consigo questões importantes sobre considerações legais e éticas que às quais também procuramos atender neste projeto [18]. A questão da factualidade, por exemplo, ou mais genericamente, da veracidade dos conteúdos, é uma das questões principais, dado que sistemas generativos podem produzir conteúdos persuasivos, mas que são incorretos, o que se pode traduzir em riscos e consequências imprevisíveis [19], [20]. Também de elevada importância, é a capacidade de rastrear a origem de conteúdos e do envolvimento humano na criação dos mesmos. A isto se acrescenta questões de autoria e de propriedade intelectual.

2.2 Oportunidade de Mercado

O crescimento e maturação rápida da inteligência artificial generativa está a levar à transformação mais significativa desta década, com investimentos projetados de centenas de milhares de milhões de euros até 2030. Este dinamismo reflete o extraordinário potencial desta tecnologia para a sociedade e o mundo. Empresas e governos estão a acelerar a sua adoção, reconhecendo que se trata de um pilar para a competitividade e também para a soberania. Por sua vez, este contexto revela uma enorme oportunidade para empresas imaginarem novos produtos e cadeias de valor.

A adoção de IA generativa tem sido transversal a quase todos os setores [21], [22], [23], [24]. Na educação, por exemplo, permite a aprendizagem personalizada; na saúde, apoia na produção de diagnósticos, na análise de documentação, e na comunicação com pacientes; na indústria, contribui para processos de design e otimização; no entretenimento, revoluciona a narração de histórias e interação com a audiência; na sociedade e administração pública, conduz à transformação digital e serviços inclusivos e centrados no cidadão.

Esta transformação assenta numa grande aceleração na produção de conteúdos digitais, mudando o paradigma de como o conhecimento e a expressão criativa são produzidos e comunicados [25], [26]. A capacidade destes sistemas de gerarem volumes enormes de material digital de elevada qualidade impulsiona a uma criação massiva e personalizada de conteúdos nos setores supra identificados, entre muitos outros. Adicionalmente, trata-se de uma tecnologia inerentemente

multilingue, capaz de construir conteúdos nas mais diversas línguas, como de os traduzir ou converter de uma para outra. Adicionalmente, estas experiências podem ser em tempo real, através da geração continua e adaptativa ao momento. A personalização será total e em tempo real, reagindo às solicitações dos utilizadores, ancorando-os à temática em questão, seja no ensino de uma matéria, ou na venda de um produto ou serviço.

3 Enquadramento

3.1 Proposta de Valor

O propósito central deste projeto consiste na investigação e desenvolvimento de uma plataforma para a construção de experiências digitais inteligentes e interativas, através do uso de inteligência artificial generativa. O objetivo é que a mesma permita a construção e publicação de experiências onde utilizadores possam participar, para as quais o sistema gera texto, imagens, áudio, e vídeo, em tempo real ou pré-preparados, ambos seguindo um programa e narrativa pré-definidos, com a finalidade de serem emotivos, interativos, personalizáveis, e adaptativos para com o utilizador.

Arquiteturalmente, pretende-se que a plataforma possua as ferramentas necessárias e a infraestrutura adequada para a configuração, produção e distribuição destas experiências. Estas experiências poderão ser decompostas em vários meios de comunicação, desde o texto até ao vídeo, como em vários formatos, desde o expositivo ao interativo e em tempo real. Em ambos os casos, a tecnologia será desenhada para permitir, sempre que possível, a sua personalização de acordo com o utilizador em questão, mediante o tipo de experiência.

Para atingir este objetivo, a plataforma irá integrar IA generativa, com modelos de interação natural, suportados por mecanismos de controlo sobre a veracidade e factualidade da experiência produzida, com a finalidade de produzir experiências coerentes com a estrutura que lhe foi pré-definida, mas também que permita à mesma de ser flexível com o utilizador, porém garantindo a sua confiabilidade.

A junção destas tecnologias, trará enorme aplicabilidade aos mais diversos setores já previamente referidos, desde a construção de módulos de formação interativos, até conversas sobre cultura em museus e locais históricos, promovendo um diálogo empático com o utilizador, e uma experiência completamente distinta da em vigor.

Com uma base tecnológica bem definida, e com a mesma a disponibilizar as ferramentas adequadas para uma construção flexível das mais diversas experiências narrativas, estas constituirão catalisadores para a inovação nos espaços criativos, culturais e educacionais. Daqui poderão surgir novos modelos de negócio, assentes sobre esta plataforma, tal como, possivelmente, novos tipos e formatos de trabalho.

Um exemplo fácil de imaginar é o percurso de um comercial ao longo do tempo. Antes, tinha de fazer prospeção porta-a-porta e marcar reuniões presenciais, porém mais tarde começou a poder realizar chamadas remotas e videochamadas, e posteriormente automatizar parte do contacto inicial com ferramentas digitais. Com esta plataforma, imaginamos o comercial a poder criar experiências de venda interativa, onde ele é representado por um agente digital treinado no seu estilo de

comunicação, argumentos e conhecimento de produto, capaz de adaptar a conversa ao perfil e necessidades de cada potencial cliente, e de ajustar exemplos, linguagem e prioridades em tempo real. Esse agente poderá também desviar-se pontualmente do guião principal para explorar oportunidades específicas que surjam durante a interação, mantendo coerência com a estratégia definida pelo comercial. Tudo previamente configurado, e ajustado dinamicamente para maximizar o impacto comercial e escalar a capacidade de contacto sem perder personalização.

Este exemplo descreve o cerne da proposta de valor do projeto. Trata-se de um sistema diferenciador, que integra a orquestração de experiências, geração de conteúdos, flexibilidade interativa, e controlo de factualidade, dentro de um produto coerente, com ferramentas disponibilizadas para o efeito.

Para realizar esta visão, o projeto foca-se em três linhas de investigação simultâneas, cada uma endereçando uma área técnica distinta, especificamente: (1) a geração controlada e multimodal de conteúdos; (2) a comunicação empática e inclusiva; (3) a orquestração de experiências. De seguida, fazemos uma breve introdução a cada uma destas.

Geração Multimodal Controlada: Foca-se em desenvolver mecanismos para o controlo semântico e estilístico na produção multimodal de conteúdos, de acordo com um objetivo, contexto e narrativa.

Diálogo Empático e Seguro: Foca-se em desenvolver agentes interativos multimodais, capazes de interagir naturalmente, tal como de percecionar as emoções do utilizador e do seu contexto circundante.

Orquestração de Experiências: Foca-se em desenvolver mecanismos para a coordenação de processos generativos dentro de ambientes dinâmicos e baseados em narrativas definidas, gerindo a progressão dos conteúdos mediante a interação do utilizador.

Consideramos que estas três linhas de investigação providenciarão as bases necessárias para a implementação bem-sucedida da plataforma idealizada.

Como guia para o desenho e implementação do sistema, foram definidos cinco pilares funcionais que o sistema deve suportar, providenciando para esse efeito um conjunto de componentes funcionais reutilizáveis, que são mais detalhados na secção **Solução Proposta**. Foi determinado que a plataforma deveria suportar funcionalmente esses cinco verticais, para servir de guia num desenho flexível das funcionalidades que a irão compor. Daqui segue também que a mesma não estará restrita a apenas esses verticais e que, devido à flexibilidade que lhe é idealizada, suportará outras aplicações ainda por idealizar. Os cinco verticais funcionais são brevemente introduzidos de seguida.

Formação Profissional: consiste em permitir a construção de simulações de formação e treino adaptativos, com interação e diálogo construtivo, para apoiar o utilizador a corresponder aos mais diversos cenários de formação, dentro de um ambiente seguro e controlado.

Educação: consiste em permitir a construção de cursos educativos para a tutoria personalizada e interativa, com apoio e diálogo pedagógico, para transmitir novo conhecimento e ajudar na aprendizagem do mesmo, semelhantemente a um professor e explicador.

Publicidade: consiste em permitir a construção de anúncios publicitários interativos, com uma narrativa que os guia, e com a capacidade de se adaptarem dinamicamente ao público que assiste ao anúncio, tornando interativo e consciente.

Cultura: consiste em permitir a construção de tours virtuais e encenações históricas, com a finalidade de emergir o utilizador dentro da história do tema ou local em questão, permitindo que o mesmo interaja com o narrador, fazendo questões e aprendendo sobre o assunto.

Cidadania: consiste em permitir a construção de meios de comunicação sociais, para o apoio ao cidadão e à sua inclusão social, respondendo-lhe a questões sobre saúde, trabalho e serviços públicos.

As linhas de investigação escolhidas e os verticais funcionais identificados promovem uma iniciativa com características estratégicas para o desenho ambicioso de uma plataforma que promove o uso da inteligência artificial generativa para a expansão da criatividade e comunicação humana, apicados à criação de novas experiências digitais interativas, com especial valor social, cultural e profissional.

3.2 Impacto Económico

Delinear e estimar o impacto económico.

- Contribuição para o reforço da economia do conhecimento
    - O projeto posiciona-se como vetor estratégico na transição para uma economia baseada em conhecimento, dados e criatividade assistida por IA.
    - Potencia o surgimento de novos produtos e serviços digitais exportáveis, criando valor acrescentado interno.
- Estímulo à modernização tecnológica de setores tradicionais (formação, educação, comunicação, cultura, serviços públicos) através da digitalização inteligente
    - Acelera a transformação digital de setores de elevado peso económico e social, introduzindo automação cognitiva e interação personalizada.
    - Na formação e educação, reduz custos e amplia alcance, permitindo formação contínua e personalizada em larga escala.
    - No setor cultural e turístico, cria novas formas de monetização e de envolvimento do público através de experiências imersivas.
    - Na comunicação e publicidade, multiplica a produtividade e permite campanhas interativas e hiperpersonalizadas.
    - Nos serviços públicos, melhora a acessibilidade, eficiência e transparência na relação com o cidadão.
    - Gera externalidades positivas sobre todo o ecossistema digital, fomentando inovação e competitividade transversal.
- Criação de emprego qualificado em áreas de engenharia de IA, design digital e conteúdos interativos
    - Contribui para a geração de emprego altamente qualificado, tanto direto (equipa de desenvolvimento) como indireto (ecossistema de inovação).
    - Estima-se a criação de postos de trabalho em engenharia de IA, ciência de dados, design interativo, UX, som e vídeo generativo, curadoria digital e ética aplicada.

4 Objetivos Científicos e Tecnológicos

4.1 Linhas Gerais de Investigação

4.1.1 Geração Multimodal Controlada

O objetivo desta linha de investigação consiste em investigar métodos para a geração coordenada de conteúdos em diferentes formatos, como texto, imagens e vídeo, com a capacidade de garantir um controlo semântico sobre a coerência e factualidade dos conteúdos, tal como o estilo da sua exposição. A mesma, é de seguida subdividida nos seguintes focos de estudo:

- Desenvolvimento de um componente para a geração multimodal controlada de conteúdos: Este foco de estudo incide na exploração de arquiteturas que permitam a desconstrução e roteamento de informações através de modelos de linguagem que atuam como controladores, coordenando a geração de conteúdos ao longo de diversos modelos secundários, de forma a obter resultados que sigam uma coerência narrativa, estilística e factual. Pode ser decomposto nos seguintes pontos:
    - Investigação de arquiteturas modulares que permitam uma orquestração flexível de modelos generativos diversos, com o objetivo de preservar a integridade contextual, coerência criativa e factualidade.
    - Investigação de mecanismos de prompt chaining e embedding de contexto, para a transferência consistente entre sequências de execução de modelos generativos.
    - Investigação de modelos híbridos, compostos por representações estruturadas de dados, e a sua transformação em conteúdos textuais e audiovisuais, gerados via IA.
- Investigação de métodos para o controlo estilístico e semântico de conteúdos: Este foco de estudo incide no desenho de mecanismos de feedback loops para a adaptação dinâmica de prompts através da observação dos outputs gerados, com a finalidade de alinhar o output de modelos de acordo com a narrativa, tom, estilo e propósito desejados. Pode ser decomposto nos seguintes pontos:
    - Definição de mecanismos de controlo que permitam a utilizadores definir narrativas, encorpadas por componentes emocionais e estilísticas, com a finalidade de garantir que o output reflita as nuances pretendidas.
    - Estudo e identificação de técnicas de feedback loops para a reestruturação de prompts com a finalidade de condicionar os modelos a seguirem os mecanismos de controlo configurados pelos utilizadores.
    - Estudo e desenho de mecanismos que permitam aos utilizadores avaliar a qualidade dos ajustes efetuados, para permitir melhoramentos iterativos.
- Investigação de métodos para a ancoragem, alinhamento e verificação factual de conteúdos: Este foco de estudo incide na experimentação de diferentes técnicas para garantir a factualidade e alinhamento de conteúdos, reduzindo o risco de desinformação e inconsistências lógicas ou sequenciais. Pode ser decomposto nos seguintes pontos:
    - Investigação de técnicas de RAG e de ancoragem para garantir a factualidade de conteúdos e prevenir alucinações e fortalecer a consistência e precisão dos conteúdos.

    - Investigação de técnicas de ancoragem para o alinhamento de vetores semânticos de conteúdos multimodais, com a finalidade de obter representações semanticamente semelhantes sobre um mesmo contexto semântico.
    - Determinação de métricas de consistência de alinhamento entre modelos multimodais.
- Investigação de métodos para a avaliação quantitativa da qualidade de modelos multimodais: Este foco de estudo incide na definição de métodos quantitativos e objetivos para a determinação da qualidade da coerência e factualidade dos conteúdos, tal como o estilo da sua exposição, de conteúdos produzidos por modelos multimodais. Pode ser decomposto nos seguintes pontos:
    - Definição de critérios que permitam mensurar a qualidade da coerência, factualidade e exposição dos conteúdos gerados por modelos multimodais
    - Construção me métodos de registo e comparação de critérios de qualidade para uma avaliação sistemática da qualidade dos mesmos

4.1.2 Comunicação e Diálogo Seguro e Empático

O objetivo desta linha de investigação consiste em investigar sobre agentes conversacionais multimodais, capazes de comunicar em tempo real de forma natural, empática e personalizada com os seus utilizadores, através do formato textual, voz e audiovisual. A mesma, é de seguida subdividida nos seguintes focos de estudo:

- Desenvolvimento de um componente para o estabelecimento de diálogos cativantes e com significado com o utilizador: Este foco de estudo incide no desenvolvimento de técnicas multimodais que providenciem interações fluídas, emotivas e dotadas de contexto com enfase na percepção e expressão multimodal que permita aos agentes interpretar e responder com uma combinação coordenada linguística, vocal e visual. Pode ser decomposto nos seguintes pontos:
    - Estudo de arquiteturas multimodais que providenciem agentes conversacionais que respondam em tempo real em diferentes formatos, de forma natural e humana.
    - Desenvolvimento de mecanismos para persistência de memória de contexto a curto, médio, e longo prazo, para garantir continuidade de fluidez ao longo de diálogos extensos.
    - Implementação de mecanismos que permitam aos agentes interpretar o contexto emotivo do utilizador, e adaptar dinamicamente o seu tom, comportamento, e complexidade de discurso.
    - Estudo da aplicação de estruturas de dados de diálogo para dar suporte à geração ancorada de respostas.
- I&D de métodos para a personalização da comunicação: Este foco de estudo incide sobre o desenvolvimento de métodos para o ajuste do tom, ritmo, complexidade do discurso, tal como a ordem e foco dos tópicos em questão, de acordo com a natureza e preferências do utilizador. Pode ser decomposto nos seguintes pontos:

    - Desenvolvimento de frameworks que recorrem a informação sobre o comportamento do utilizador e o contexto em que se insere, para ajustar a sua linguagem, tom e complexidade de discurso.
    - Estudo de métodos de medição do envolvimento do utilizador através da quantificação de variáveis de interesse, como conteúdo semântico das respostas e tempos de atenção, para fornecer informações úteis para métodos de ajuste comportamental do agente.
- I&D de métodos de análise emocional e para o estabelecimento de empatia: Este foco de estudo incide em identificar técnicas que dotem os agentes de inteligência emocional, com a finalidade de os encaminhar a responder de forma adequada e contextualizada para os estados emotivos do utilizador. Pode ser decomposto nos seguintes pontos:
    - Investigar modelos que permitam identificar estados emocionais através da análise de áudio e de vídeo do utilizador.
    - Desenvolver mecanismos de classificação coerente ao longo do tempo do estado emocional do utilizador mediante os valores detetados pelos modelos utilizados.
    - Definição de métricas para a quantificação da situação emocional do utilizador e da capacidade do sistema de transmitir empatia e fortalecer o relacionamento com o utilizador.

4.1.3 Orquestração de Experiências e Narrativas Interativas

O objetivo desta linha de investigação consiste em investigar métodos para a construção de experiências narrativas, que permitam que se desviem temporariamente da linha narrativa mediante as interações do utilizador e o contexto em que está inserido. A mesma, é de seguida subdividida nos seguintes focos de estudo:

- Desenvolvimento de componente para a orquestração de narrativas: Este foco de estudo incide na investigação de uma ferramenta para a composição de narrativas, de forma a entregar informações estruturais e guias de execução a modelos generativos, com a finalidade de os guiar sobre que conteúdos gerarem mediante o contexto da narrativa e o contexto do utilizador. Pode ser decomposto nos seguintes pontos:
    - Desenho de ferramenta de composição de narrativas, para serem geradas tanto em tempo real como de forma antecipada, permitindo ao utilizador uma fácil composição da história, com a orquestração de elementos expositivos e interativos, tal como a flexibilidade que os mesmos têm em se adaptar ao utilizador.
    - Desenho de ferramenta para a configuração dos espaços e limitações que elementos narrativos oferecem ao utilizador para os manipular ou alterar mediante as suas interações.
- I&D de métodos para garantir uma continuidade narrativa: Este foco de estudo incide em identificar técnicas para garantir a consistência lógica e temporal dentro de uma dada experiência, mediante as interações e progresso do utilizador. Pode ser decomposto nos seguintes pontos:
    - Estudar mecanismos de memória narrativa para o registo de eventos, progressão e decisões do utilizador.

- Estudar mecanismos que determinam a consistência causal para guiar e alinhar modelos generativos a produzirem conteúdos coerentes e factuais.
- Estudar ferramentas de autor que auxiliem os designers de narrativas a focarem-se nos elementos importantes e a delegarem os detalhes aos modelos, promovendo uma fronteira produtiva entre os dois.

4.2 Implementação e Produtização

Descrever como as linhas de investigação dão suporte à implementação de um protótipo, enquadrar também com os verticais funcionais, e explicar como é que o mesmo poderá depois ser produtizado.

- Desenvolvimento de protótipo que combine os resultados das três linhas de investigação
- Enquadramento funcional nos verticais:
  - Formação e Treino Profissional
  - Ensino e Educação
  - Publicidade e Comunicação Interativa
  - Cultura e Turismo – storytelling adaptativo e visitas guiadas personalizadas
  - Saúde, Cidadania e Serviços Públicos – assistentes informativos, empáticos e acessíveis
- Validação experimental com pilotos demonstradores para cada vertical
- Produtização: definir modelos de negócio baseados em SaaS/API, para comercialização modular das tecnologias desenvolvidas

4.3 Resultados Esperados

Indicar os principais resultados científicos, tecnológicos e sociais a alcançar no final do projeto.

5 Estado de Arte

5.1 Modelos Multimodais

O projeto em análise tem por base um uso intensivo de modelos multimodais de IA generativa, pelo que é importante rever a composição interna dos mesmos. Atualmente, a construção destes modelos segue uma de duas abordagens possíveis, (1) uma arquitetura multimodal unificada, treinada para processar e gerar múltiplos tipos de formatos pela mesma rede, tipicamente uma rede transformer e (2) uma arquitetura constituída por múltiplos modelos especialistas, tipicamente com a sua utilização coordenada por um modelo central com o qual é feita a comunicação [27], [28], [29].

São duas abordagens distintas, com as suas respetivas vantagens e desvantagens. No caso da abordagem centralizada, o processamento e geração tiram partido de uma representação interna centralizada, o que promove uma melhor coerência entre formatos de conteúdos. Em contrapartida, modelos separados beneficiam de terem toda a sua construção neural otimizada

para um dado formato de output. A escolha entre uma ou outra abordagem não é trivial, dependendo muito do contexto de utilização. Fatores a considerar, entre outros, serão: complexidade do problema, tempo de treino, qualidade requerida para o output e tempo de latência.

Para uma interligação entre diferentes formatos dentro de um modelo ou entre diferentes modelos especialistas, cada um com o seu formato, é necessário um espaço semântico partilhado para o respetivo alinhamento intermodal. Este conceito foi inicialmente apresentado pela OpenAI em 2021, no estudo em [30] que demonstra a capacidade de alinhamento multimodal através do mapeamento de texto e imagens para um mesmo espaço vetorial. Este conceito foi depois expandido para mais domínios, como demonstrado em [31] com imagem, texto e áudio, entre outros, todos representados dentro do mesmo espaço latente. Em [32] pode também ser encontrado um levantamento mais abrangente deste tipo de métodos.

A popularidade de chatbots inteligentes, aliada ao surgimento de arquiteturas multimodais levou à subsequente construção de modelos de interação por voz, e mais recentemente por vídeo e voz, os quais pedem por infraestrutura de ainda maior dimensão e especializadas, para conseguirem atingir latências baixas. Interessantemente, os modelos de visuais permitem um ajuste explícito sobre o conteúdo gerado, através de guias estruturais que encaminham a geração a incorporar a estrutura, tal como demonstrado no estudo ControlNet em [33], como em outros subsequentes [34], [35], [36]. Semelhantemente, na voz também é possível encaminhar o timbre e emoção a partir de exemplos de referência, tal como demonstrado em [37], [38], [39].

Ainda que tenhamos várias formas de encaminhar estes modelos a gerarem o seu conteúdo de acordo com um conjunto de diretrizes, todos os modelos sofrem de alucinações, i.e. de inventarem conteúdos que não são desejados ou verdadeiros, podendo estes aparecer em qualquer um dos formatos de saída. Há várias áreas de investigação para atacar este problema, incluindo o estudo de métodos de deteção e medição de alucinações [40], de pré-correção durante o treino [41], de mitigação durante a geração [42] e de correção posterior à geração [43]. No entanto, no caso de uso de IA generativa como um serviço, as técnicas de mitigação de alucinações são diferentes, focando-se em técnicas de ancoramento [44], na validação da saída do modelo [45], [46] e na revisão/atualização das prompts [47].

5.2 Memória e Personalização

A estratégia para dotar um modelo de memória – um contexto adicional ao histórico de conversa com o utilizador – tem por base técnicas de recuperação de conhecimento, denominadas por Retrieval-Augmented Generation. Estas consistem em consultar informação externa, seja de bases de dados, na internet, ou em documentação, para acrescentar conteúdo ao contexto da conversa e promover respostas mais contextualizadas do modelo. Métodos mais avançados de RAG podem consistir em técnicas encadeadas de pesquisas, onde entre cada pesquisa o modelo decide se continua mediante os resultados que obteve [48], [49].

As técnicas de pesquisa também têm evoluído nos últimos anos. Inicialmente, contava-se com pesquisas por palavras-chave ou pesquisas na internet. Mais recentemente, começou-se a utilizar intensivamente pesquisas semânticas. Estas consistem em armazenar os conteúdos em bases de dados de vetores, que por sua vez permitem encontrar resultados que são semanticamente

parecidos com a pesquisa que o utilizador está a fazer e não textualmente iguais, e.g. caractere a caractere. Em [50], é apresentada uma técnica avançada para a construção de memória a longo prazo de um agente. Esta consiste na utilização de uma base de dados de grafos para armazenar notas estruturadas com descrições contextuais, e com ligações para outras notas relacionadas. O objetivo desta abordagem consiste em construir um grafo de conhecimento sobre os mais diversos temas. Outras técnicas de gestão de memória podem ser encontradas no levantamento feito em [51].

5.3 Computação Empática

Estes sistemas interativos necessitam cada vez mais da capacidade de reconhecer e reagir adequadamente às emoções humanas. A interação pessoa-máquina está a evoluir para ambientes mais naturais, onde a componente emocional é uma parte central da comunicação. Aqui, abordagens multimodais também são importantes, dado a abrangência de meios e riqueza de formas com que os humanos se expressam, muitas vezes com elevada nuance, tanto verbal, como facial e postural. Sinais paralinguísticos, micro-expressões, variações de tom de voz e dinâmica corporal fornecem pistas complementares que dificilmente podem ser capturadas por um canal isolado. Estes modelos são denominados de modelos de reconhecimento de emoção multimodal (MER). Em [52] é feito uma apresentação exaustiva destes modelos, descrevendo as técnicas principais de fusão de múltiplos formatos (texto, voz, vídeo) para um espaço latente partilhado, onde cada um destes formatos é pré-processado com uma técnica respetiva, nomeadamente embeddings para texto, espectrogramas para áudio e CNNs para imagem. A construção de espaços latentes comuns permite que o modelo aprenda relações intrínsecas entre modalidades, preservando complementaridade e reduzindo redundâncias, melhorando a precisão emocional. Estas fusões podem depois ser analisadas também relativamente à sua sequência, para entender emoções mais complexas, que requerem algum contexto para serem entendidas, como o sarcasmo ou a ironia. O tratamento da dimensão temporal torna-se importante, pois emoções emergem em fluxo e dependem de memória contextual e continuidade expressiva. Em [53] também é feito um levantamento exaustivo de estudos, com a finalidade de demonstrar sem margem para dúvidas que a deteção emocional através de técnicas multimodais é superior à unimodal.

O reconhecimento da emoção é meio caminho para a computação empática. A segunda parte consiste em efetivamente responder de forma adequada às emoções demonstradas pelo utilizador. Para isso, os modelos são treinados para responder de forma empática e afetiva. Abordagens modernas incluem o treino com diálogos empáticos. Há vários datasets compostos para o efeito [54], [55], [56]. Além da empatia, é depois importante também haver um alinhamento cultural com o utilizador. A emoção não é universal, mas sim influenciada por normas sociais e expressões regionais. Em [57] é observado que existe um viés ocidental, derivado do corpus de dados com que os modelos são inadvertidamente treinados. Isto pode resultar em respostas que soam naturais para utilizadores ocidentais, mas estranhas ou até ofensivas noutras culturas. Em [58] é referido que existe um viés relativamente à definição de beleza e atratividade e em [59] é demonstrado que existe um viés cultural relativamente a expressões idiomáticas, vestuário e gestos. Para complicar a situação, em [60] é explicado como contextos com significados culturais distintos pode levar os modelos a confundirem-se e a estereotiparem a situação.

5.4 Orquestração de Narrativas

Estes modelos generativos permitem a criação de histórias e narrativas com interatividade. Um grande desafio, porém, consiste em garantir a sequência lógica, causal e temporal da narrativa experienciada, dado que estes modelos alucinam e podem confundir-se com a ordem cronológica de eventos e as consequências diretas ou indiretas dos mesmos. Sem o recurso a métodos de ancoragem, estes modelos ao gerarem histórias longas podem começar a contradizer-se e até entrar em ciclos dedutivos erróneos. Estas situações prejudicam a imersão do utilizador, particularmente em ambientes interativos onde escolhas passadas deveriam condicionar eventos futuros, caso contrário a ilusão de continuidade narrativa colapsa.

A utilização de RAG e de estruturas narrativas é importante, e estudos recentes propõem abordagens híbridas, onde parte da narrativa está estruturada e outra parte é entregue ao modelo para complementar. Este paradigma tem-se mostrado eficaz na redução de contradições, pois cria um esqueleto narrativo que funciona como guia e âncora conceptual. A componente RAG acrescenta uma camada factual e contextual, permitindo que eventos prévios, descrições de personagens e regras do mundo sejam recuperados. Esta abordagem permite também que a resposta do modelo se enquadre melhor com o contexto do utilizador. A base estrutural consiste tipicamente num conjunto de capítulos e pontos de salto ou de retrocesso, que encaminham o modelo na direção certa, deixando-lhe espaço para compor os detalhes [61]. Em [62] é demonstrado que esta abordagem facilita a construção coerente de histórias longas. O modelo opera guiado por um mapa narrativo, preenchendo os detalhes, mas condicionado por checkpoints estruturais. O resultado tende a apresentar melhor consistência causal, manutenção de temas e evolução plausível de personagens. Na componente visual, em [63] são aplicadas técnicas de memória visual explícita para garantir que elementos visuais se mantenham constantes durante a história.

Técnicas semelhantes são exploradas em [64], articulando geração narrativa com imagens coerentes ao longo do tempo, orquestrando técnicas (busca, difusão, edição) para preservar identidade, tom e continuidade estética, e em [65] este tipo de técnicas são usadas para a consistência de personagens em conteúdos Manga. Estas abordagens permitem criar mundos narrativos que evoluem de forma consistente, um fluxo narrativo coerente e integrado, em que texto e imagem suportam e reforçam mutuamente a construção de sentido, coerência e imersão.

6 Solução Proposta

6.1 Componentes Funcionais

6.1.1 Conhecimento e Governança

6.1.1.1 Gestão de Utilizadores & Perfis

Sistema de perfis, preferências e histórico que suporta personalização, controlo de acesso e consentimentos de dados.

Funcionalidades:

A aplicação permite o registo e login seguros através de email/palavra-passe ou SSO. Os utilizadores podem gerir as suas preferências, consultar e alterar as suas informações pessoais, bem como acompanhar o histórico de atividades após efetuarem login. Comprometemo-nos a proteger os seus dados e sessões, em conformidade com as normas de privacidade.

- **Registo e Login**
  E possível o registo e login utilizando email/palavra-passe ou SSO (com MFA e verificação de email). Garantindo também a recuperação de acesso à conta, estando em conformidade com o RGPD.

- **Gestão de Preferências e Consentimentos**
  Após o login, os utilizadores podem alterar as suas preferências e consentimentos de forma detalhada, como desativar notificações e partilhar dados, podendo fazê-lo a qualquer momento.

- **Consulta e Alteração de Dados Pessoais**
  Os utilizadores podem ver e editar o seu perfil na área pessoal.

- **Consulta de Histórico**
  É possível consultar o histórico de sessões, interações e conteúdo gerado, facilitando o acompanhamento das ações realizadas na aplicação.

- **Sessão de Utilizador**
  Tokens de curta duração mantêm a sessão do utilizador ativa, existindo também mecanismos de proteção CSRF para reforçar a segurança.

Dependências: -

6.1.1.2 Segurança, Ética e Conformidade

Segurança e conformidade com o RGPD, AI Act e princípios éticos de transparência, neutralidade, segurança e filtragem de conteúdos inadequados.

Funcionalidades:

A aplicação implementa padrões de segurança, ética e conformidade que se traduzem em transparência e proteção dos dados. Inclui funcionalidades que permitem gerir políticas e consentimentos, controlar a aprovação editorial e ética, e garantir a rastreabilidade através de logs de auditoria.

- **Gestão de políticas ativas e consentimentos**
  Permite administrar políticas de privacidade, segurança e uso de dados, bem como gerir os consentimentos dos utilizadores de forma centralizada e transparente. Garante que todas as definições estejam alinhadas com os requisitos legais e éticos aplicáveis. É possível atualizar e revogar os consentimentos a qualquer momento.

- **Workflow de aprovação editorial e ética**
  Inclui processos de revisão e aprovação de conteúdos, de forma que todas as publicações respeitam as normas editoriais, princípios éticos e regulamentos em vigor. O workflow permite que diferentes utilizadores participem na validação dos conteúdos antes da sua divulgação. Ajuda a prevenir a publicação de informação inadequada ou sensível.

- **Logs de auditoria e mecanismos de rastreabilidade**
  Regista todas as ações relevantes dos utilizadores na plataforma, sendo possível monitorizar alterações, decisões e acessos a dados. Transparência e responsabilidade é garantida através dos logs de auditoria, facilitando a identificação de potenciais irregularidades. Estes mecanismos ajudam a garantir o cumprimento das normas de segurança e de proteção de dados.

Dependências: -

6.1.1.3 Gestão de Conhecimento

Repositório central para organização e versionamento de conteúdos, modelos e dados contextuais.

Funcionalidades:

A utilização de um repositório centralizado permite a organização e a gestão durante todo o ciclo de vida do conteúdo, dos modelos e dos dados contextuais, apoiando a partilha de conhecimento e a melhoria contínua.

- **Ingestão documental**
  Organiza a importação de documentos/recursos para o repositório. Permite a organização com atribuição de metadados para uma categorização eficaz. Garante a consistência e a rastreabilidade de todos os materiais ingeridos.

- **Armazenamento e pesquisa de modelos, prompts e recursos multimodais**
  Ajuda a manter o controlo de todos os materiais ingeridos de forma consistente para posterior recuperação através da recolha e rastreabilidade. Facilita o armazenamento seguro e a recuperação rápida de modelos, prompts e dados multimédia. Uma pesquisa avançada ajuda os utilizadores a encontrar rapidamente os recursos apropriados.

- **Versionamento e ciclo de vida (Rascunho → Revisão → Aprovado → Publicado → Descontinuado)**
  Rastreia o estado de cada item desde a sua criação até à sua desativação e mantém as atualizações e aprovações sob controlo. Mantém um registo transparente das alterações e da propriedade ao longo de todo o processo. Cumpre os requisitos de conformidade e governança.

- **Gestão de dados**

Apoia a organização, manutenção e integridade de diferentes tipos de dados no repositório. Estabelece controlos de acesso para proteger dados sensíveis. Facilita a análise e a geração de relatórios para fundamentar a tomada de decisões orientada por dados.

Dependências: -

6.1.1.4 Biblioteca de Conhecimento

Ferramentas para a organização de catálogos de conhecimento, da construção de blocos reutilizáveis de atividades, templates, e estruturas narrativas. Integra com a **Gestão de Conhecimento**.

Funcionalidades:

A Biblioteca de Conhecimento inclui ferramentas para organizar catálogos de conhecimento e desenvolver blocos de atividades reutilizáveis, modelos e estruturas narrativas, que se integram na Gestão do Conhecimento.

- **Organização de conteúdos / templates / blocos / catálogos / estruturas de atividades / storytelling reutilizáveis**
  Permite a organização estruturada de diversos recursos, como modelos, blocos de conteúdo e estruturas narrativas facilitando o acesso e a adaptação de diversos materiais de forma a atender a diferentes necessidades de aprendizagem e criatividade.

- **Pesquisa de recursos**
  Disponibiliza uma pesquisa avançada para localizar rapidamente modelos, blocos e itens de catálogo relevantes garante que os utilizadores possam encontrar e utilizar eficientemente os materiais certos para os seus projetos.
  O suporte de design modular para atividades e elementos de narrativa, melhora a consistência, a eficiência, a descoberta e a recuperação de recursos na biblioteca.

Dependências:

- ***Gestão de Conhecimento***

6.1.1.5 Validação com Grounding

Indexação de fontes fiáveis para o apoio à geração e verificação da factualidade de conteúdos gerados.

Funcionalidades:

A validação de fundamentos é concebida para apoiar a geração e verificação de conteúdo factual através da indexação de fontes fidedignas, garantindo maior precisão e credibilidade.

- **Validação factual e semântica de conteúdos**
  O conteúdo é validado em relação a fontes fidedignas indexadas quanto à veracidade e integridade semântica. Ao garantir a qualidade da informação, este processo minimiza as oportunidades para declarações falsas e gera resultados contextualmente apropriados e fiáveis.

- **Atribuição de scores de veracidade e alinhamento**
  Cada conteúdo gerado recebe uma pontuação com base na sua correção factual e alinhamento com as fontes validadas. Estas pontuações não são apenas transparentes, como também permitem aos utilizadores determinarem facilmente a fiabilidade e a credibilidade das informações fornecidas.

- **Deteção de inconsistências ou contradições**
  O conteúdo é analisado por ferramentas buscam automaticamente inconsistências ou contradições lógicas nas declarações em comparação com as referências indexadas. Quando estas discrepâncias são detetadas, desencadeiam alertas para garantir a reavaliação, correção ou outras medidas que apoiem a coerência e a credibilidade global da base de conhecimento.

Dependências:

- ***Segurança, Ética e Conformidade***
- ***Gestão de Conhecimento***

6.1.2 Geração de Conteúdos

6.1.2.1 Motor de Geração Multimodal

Sistema generativo capaz de produzir texto, imagem, voz e vídeo de forma coordenada e coerente com o contexto e estilo solicitados.

Funcionalidades:

O motor de geração de conteúdo é um sistema multimodal desenhado para produzir texto, imagens, áudio e vídeo de uma maneira contextualmente coordenada e alinhada com os requisitos e estilo solicitado.

- **Geração de conteúdos multimodais**
  É disponível a criação de diversos tipos de matérias como texto escrito, recursos visuais, áudio e vídeo e ao tirar proveito de modelos generativos avançados é assegurado que os conteúdos são contextualmente relevantes e podem ser perfeitamente integrados de forma a criar experiências mais ricas para o utilizador.

- **Ajuste de estilo, tom e coerência**
  A geração de conteúdo pode ser ajustada para corresponder especificamente a as preferências estilísticas e tonais e ao mesmo tempo manter a coerência. Os utilizadores podem especificar as características e o sistema de forma inteligente e consistente adapta o conteúdo gerado para corresponder ao contexto, audiência e objetivos de comunicação.

- **Aplicação de guardrails de segurança e ética**
  Durante todo o processo de geração existe um processo de segurança onde são aplicadas barreiras éticas. Estas medidas de segurança que seguem normas de utilização

responsável, monotorização o conteúdo para riscos potencias garantindo um uso responsável e em conformidade com as orientações legais e éticas. De forma as proteger os utilizadores o conteúdo inapropriado e/ou maldoso é filtrado e sinalizado.

Dependências:

- ***Validação com Grounding***

6.1.2.2 Apoio Construção Multilingue

Tradução automática e apoio à adaptação cultural dos conteúdos a serem produzidos pela **Autoria Inteligente** tal como pelo **Agente Conversacional**.

Funcionalidades:

O suporte multilíngue oferece traduções automáticas e adaptação cultural ao conteúdo gerado melhorando a acessibilidade e relevâncias para diversos tipos de público-alvo.

- **Legendagem de conteúdos de vídeo**
  Geração automaticamente de legendas para conteúdos de vídeo com suporte de múltiplos idiomas. Desta forma é garantido um maior envolvimento e compreensão, pois os conteúdos multimédia são acessíveis a todos independente do seu idioma.

- **Tradução integrada de conteúdos**
  Através de modelos de linguagem avançados existe uma tradução integrada de texto, áudio e conteúdos visuais envolvente a todo o sistema. Esta tradução preserva o significado e contexto ao longo do processo de criação de forma a possibilitar os utilizadores a produzir conteúdos facilmente compreendidos em todo o mundo.

- Validação e orientação para adequação cultural
  De forma a oferecer recomendações e melhorias, a relevância e a sensibilidade cultural do conteúdo traduzido é avaliada. A aplicação identifica potencias problemas e oferece orientações para garantir que os conteúdos são respeitosos e relevantes para o público-alvo.

Dependências:

- ***Biblioteca de Conhecimento***
- ***Motor de Geração Multimodal***

6.1.2.3 Autoria Inteligente

Ferramenta de criação assistida por IA para conceber conteúdos multimodais de forma rápida e personalizada.

Funcionalidades:

A Autoria Inteligente com suporte a inteligência artificial permite aos utilizadores a rápida criação personalizada de conteúdos multimodais, e oferece capacidades avançadas de organização, verificação e suporte multilingue.

- **Geração de conteúdos multimodais**
  Permite aos utilizadores criar rapidamente conteúdo personalizado de texto, imagens, áudio e vídeo para diferentes necessidades. O sistema garante que os dados são gerados de forma coerente, contextualizados e criativamente apropriados.

- **Organização de conteúdos dentro da biblioteca**
  Permite o armazenamento a organização estruturada da informação dentro da biblioteca. Útil para indexação, recuperação e gestão de recursos, tanto para acessibilidade como para reutilização a longo prazo.

- **Apoio assistido por IA à criação e organização dos conteúdos**
  Utiliza a inteligência artificial para oferecer conselhos aos utilizadores que precisam de criar ou estruturar conteúdo. Desta forma existe um fluxo de trabalho mais eficiente e um aumento da qualidade e a eficácia na produção de conteúdos.

- **Verificação factual e ética durante autoria**
  Durante a criação de conteúdos existe uma verificação da informação e da conformidade ética em tempo real. São despoletados alertas sobre inconsistências ou riscos, de forma a gerar conteúdos fiáveis e responsáveis.

- **Suporte multilingue integrado**
  Oferece tradução automática e adaptação cultural, permitindo a criação de conteúdos para públicos diversificados. Isto alarga o alcance dos materiais gerados, pois o conteúdo é acessível em diferentes idiomas e regiões.

Dependências:

- ***Biblioteca de Conhecimento***
- ***Motor de Geração Multimodal***
- ***Apoio Construção Multilingue***

6.1.3 Experiências Não-Interativas

6.1.3.1 Apresentação de Conteúdos

Gere a apresentação de conteúdos já existentes (ex: cursos, módulos, tours, campanhas) de acordo com perfis, permissões e contexto de utilização.

Funcionalidades:

As experiências não interativas apresentam um formato estruturado de apresentação de conteúdos, como cursos e módulos, de acordo com os perfis, permissões e contexto de utilização dos utilizadores

- **Coordena a apresentação de conteúdos da biblioteca de conhecimento**
  Controla a forma como o conteúdo da biblioteca de conhecimento é apresentado e altera a seleção com base nas funções e no contexto do utilizador. Ao apresentar aos utilizadores apenas o que é necessário ou exigido, garante que estes têm o material necessário para aprender e interagir mais eficazmente.

- **Gestão de publicação, entitlements e versões ativas**
  Supervisiona o ciclo de vida da publicação do conteúdo publicado, incluindo os direitos e o controlo de versões. O acesso dos utilizadores depende das suas permissões e apenas têm acesso a versões atualizadas e aprovadas, mantendo a coerência e a conformidade.

- **Pesquisa, filtragem e personalização de catálogos**
  Permite a navegação e a filtragem de catálogos de conteúdos para facilitar a descoberta de material relevante. As funcionalidades de personalização adaptam a visualização do catálogo ao perfil e às preferências dos utilizadores, criando assim uma experiência de utilizador mais personalizada e relevante.

Dependências:

- ***Biblioteca de Conhecimento***

6.1.4 Experiências Interativas

6.1.4.1 Perceção Emocional e Empatia

Analisa emoções e tom do utilizador, ajustando a resposta do sistema para promover interações empáticas.

Funcionalidades:

Criação de uma experiência interativa onde a interpretação das emoções e intenções dos utilizadores são capturadas e analisadas possibilitando responder-lhes com respostas empáticas e personalizadas, e assim aumentar o seu nível de envolvimento e satisfação.

- **Leitura de emoções físicas**
  Usando visão computacional a aplicação executa uma análise física como expressões faciais, gestos e postura. Isto garante a identificação do estado emocional do utilizador de forma que possa ser dada uma resposta sensível e adequada.
- **Leitura de tom de voz**
  Ao avaliar o tom, a entoação e a inflexão da voz do utilizador o sistema deteta emoções e o estado de espírito do utilizador. Esta perceção é utilizada para ajustar as respostas garantindo que as interações estão emocionalmente sintonizadas e dentro do contexto correto.

- **Leitura de intenções do utilizador**

Técnicas de linguagem natural e analise de comportamento são usadas para interpretar as intenções dos utilizadores durante as interações. Desta forma é possível antecipar necessitadas, fazer ligações relevantes e garantir uma conversa mais empática e compreensiva.

Dependências: -

6.1.4.2 Orquestrador de Experiências

Gere o fluxo e a lógica das experiências interativas, coordenando simulações, conteúdos, decisões e narrativa em tempo real.

Funcionalidades:

O Orquestrador de Experiências como responsável pela gestão do fluxo e lógica das experiências interativas a sua função é coordenar as simulações, conteúdo, decisões e narrativa em tempo real de forma a criar uma jornada de utilizador envolvente e personalizada.

- **Planeamento, Composição e Coordenação do Fluxo Narrativo**
  O orquestrador cria e gere o fluxo narrativo que chega aos utilizadores. A sua função principal é adaptar a narrativa baseando-se no contexto, na experiência adquirida, e no estado emocional do utilizador. Desta forma garante que cada interação têm um maior impacto emocional e relevância melhorando assim a imersão e o envolvimento em toda a experiência.

- **Monitoriza a progressão do utilizador e atualiza o seu plano e fluxo**
  Monitoriza a jornada e o progresso do utilizador dentro da experiência interativa de forma a ajustar o fluxo da narrativa e ações futuras. Ao atualizar dinâmica os planos e conteúdo, o sistema de forma transparente alinha-se com as necessidades e preferências do utilizador.

Dependências:

- ***Biblioteca de Conhecimento***
- ***Perceção Emocional e Empatia***

6.1.4.3 Agente Conversacional Multimodal

Agente de diálogo em tempo real, que integra texto, voz e visão, permitindo uma comunicação natural e contextual com o utilizador.

Funcionalidades:

O Agente Conversacional Multimodal é um agente de comunicação que combina texto, voz e visuais em tempo real para garantir uma comunicação natural e contextualizado com os utilizadores.

- **Gera conteúdos multimodais em tempo real**

O agente cria conteúdo textual, áudio e visual instantaneamente, ajustando os resultados com base na conversa em curso e ao contexto do utilizador. Isto permite uma integração de várias modalidades, criando uma experiência mais rica e imersiva para os utilizadores durante a interação.

- **Escuta o utilizador e interrompe a geração se necessário**
  O agente escuta ativamente o utilizador durante todo o diálogo, pausando a produção de conteúdo caso o utilizador interrompa. Esta é uma forma fácil de obter um controlo conversacional mais orgânico e fazer com que as interações fluam de forma mais natural.

- **Armazena histórico de interações com o utilizador por cada experiência**
  Regista cada interação com o utilizador por experiência. Este histórico oferece sugestões de respostas personalizadas, melhorando continuamente a compreensão contextual em futuras interações garantindo uma experiência de utilizador mais rica e envolvente.

Dependências:

- ***Biblioteca de Conhecimento***
- ***Motor de Geração Multimodal***
- ***Orquestrador de Experiências***

6.1.5 Acesso e Entrega

6.1.5.1 Acesso e Comunicação

Interfaces web para a obtenção e transmissão dos conteúdos aplicacionais e dos conteúdos multimodais, para cada utilizador.

Funcionalidades:

A aplicação adapta os conteúdos e funcionalidades de acordo com o perfil e permissões do utilizador, garantindo segurança e personalização. Integrações robustas com APIs, streaming e serviços externos promovem interoperabilidade e uma experiência eficiente e fluida.

- **Disponibiliza as funcionalidades e conteúdos da aplicação mediante o utilizador e permissões de acesso**
  Com base em quem é o utilizador e o seu tipo acesso, apenas são apresentados certos conteúdos e funcionalidades. Desta forma é garantindo que o utilizador apenas vê informação e ferramentas relevantes e ao mesmo tempo é garantido uma melhoria da segurança, personalização e fluxo de trabalho dentro da aplicação

- **Gestão de APIs, streaming e integrações externas**
  Ferramentas de gestão robustas permitem a integração de serviços externos através de API, streaming de conteúdos em tempo real e conectividade com plataformas de terceiros. Garantindo uma interoperabilidade e escalabilidade de forma a facilitar a experiência do utilizador e uma troca de dados fluida.

Dependências:

- ***Gestão de Utilizadores & Perfis***
- ***Segurança, Ética e Conformidade***
- ***Biblioteca de Conhecimento***
- ***Autoria Inteligente***
- ***Agente Conversacional Multimodal***

6.1.5.2 Frontend Aplicacional

Camada de apresentação web multimodal, com suporte para diferentes formatos alternativos, como mobile, quiosque e VR, que disponibiliza as experiências ao utilizador final.

Funcionalidades:

A interface da aplicação oferece uma camada de apresentação web multimodal, através de vários formatos de apresentação, como dispositivos móveis, quiosques e realidade virtual, permitindo aos utilizadores uma experiência personalizada aos conteúdos interativos.

- **Disponibiliza visualmente as funcionalidades e conteúdos da aplicação mediante o utilizador e permissões de acesso**
  A interface apresenta os recursos e informações da aplicação de acordo com o perfil e as permissões do utilizador. Isto significa que os utilizadores apenas têm acesso ao que é relevante e às ferramentas disponíveis, o que aumenta a segurança e a personalização da experiência

- **Captura de interações**
  As interações do utilizador são monitorizadas, independentemente do formato ou dispositivo utilizado. Isto permite que existam adaptações em tempo real e feedback personalizado garantindo melhorias contínuas na interface do utilizador e na experiência geral.

Dependências:

- ***Acesso e Comunicação***

6.1.6 Observabilidade e Melhoria

6.1.6.1 Analítica

Recolha e análise de dados sobre interação, tempo de utilização e eficácia de experiências.

6.1.6.2 Melhoria Adaptativa

Transforma dados analíticos em sugestões de melhoria, fornecendo recomendações para o melhoramento da plataforma.

6.2 Verticais Funcionais

6.2.1 Formação e Treino Profissional

Exemplos: Programas de onboarding automatizados e interativos; Treino de vendas, compliance, segurança ou operação de equipamentos; Simulações de clientes para equipas comerciais ou de suporte.

6.2.2 Ensino e Educação

Exemplos: Criação automática de módulos educativos multimodais (vídeo, áudio, texto, interação); Aulas adaptativas com feedback instantâneo; Recursos acessíveis e inclusivos para alunos com diferentes necessidades.

6.2.3 Publicidade e Comunicação Interativa

Exemplos: Vídeos interativos que respondem a perguntas do cliente; Demonstrações virtuais de produtos, com personalização por perfil de utilizador; Continuidade de marca entre publicidade e pós-venda (assistente que mantém a "voz" da marca).

6.2.4 Cultura, Turismo e Património

Exemplos: Guias virtuais em museus e locais históricos com narrativa adaptativa; Recriação de personagens históricas para fins educativos e turísticos; Experiências imersivas em exposições, feiras e eventos culturais.

6.2.5 Saúde, Cidadania e Serviços Públicos

Exemplos: Educação de pacientes e formação de cuidadores; Assistentes digitais que explicam serviços públicos e direitos do cidadão; Campanhas de sensibilização personalizadas (saúde, ambiente, segurança).

6.3 Arquitetura Geral

7 Plano de Execução

7.1 Abordagem Metodológica e Experimental

Definir a metodologia científica, fases de desenvolvimento e validação do projeto.

7.2 Cronograma Geral

Apresentar o planeamento temporal, os marcos principais e a sequência das atividades de I&D.

7.2.1 Work Packages

7.2.1.1 WP1

7.2.2 Planeamento

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
WP1																								

8 Referências

Outras referências – não estão incluídas no texto mas queremos que apareçam na lista de referências: [13], [62], [63], [64], [65], [66], [67], [68]

[1] "Agentifai (Alice) — Assistente Virtual com IA Generativa," 2025.

[2] "Visor.ai — Plataforma de Automação de Atendimento com IA," 2025.

[3] "Pitch Avatar (ROI4Presenter) — Apresentações Interativas com Avatar IA," 2025.

[4] "D-ID — Avatares Interativos e Vídeo Gerado por IA," 2025.

[5] "UneeQ — Digital Humans for Customer Experience," 2025.

[6] "Synthesia — Criação de Vídeo com IA e Avatares Virtuais," 2025.

[7] "HeyGen — Criação de Vídeo e Avatares Interativos com IA," 2025.

[8] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," May 27, 2015, *Nature Publishing Group*. doi: 10.1038/nature14539.

[9] A. Vaswani *et al.*, "Attention is all you need," *Adv Neural Inf Process Syst*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.

[10] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.02155

[11] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," 2020.

[12] OpenAI *et al.*, "GPT-4 Technical Report," Mar. 2023, [Online]. Available: http://arxiv.org/abs/2303.08774

[13] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan, "Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models," Mar. 2023, [Online]. Available: http://arxiv.org/abs/2303.04671

[14] "Hugging Face – The AI community building the future." Accessed: Nov. 05, 2025. [Online]. Available: https://huggingface.co/

[15] N. Liu, L. Chen, X. Tian, W. Zou, K. Chen, and M. Cui, "From LLM to Conversational Agent: A Memory Enhanced Architecture with Fine-Tuning of Large Language Models," Jan. 2024, [Online]. Available: http://arxiv.org/abs/2401.02777

[16] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," in *UIST 2023 - Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*,

Association for Computing Machinery, Inc, Oct. 2023. doi: 10.1145/3586183.3606763.

[17] P. Office of the European Union L- and L. Luxembourg, “Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)Text with EEA relevance.” [Online]. Available: http://data.europa.eu/eli/reg/2024/1689/oj

[18] Z. Ji *et al.*, “Survey of Hallucination in Natural Language Generation,” Jul. 2024, doi: 10.1145/3571730.

[19] Y. Wang *et al.*, “Factuality of Large Language Models: A Survey,” Oct. 2024, [Online]. Available: http://arxiv.org/abs/2402.02420

[20] Fengchun. Miao, Stefania. Giannini, and Wayne. Holmes, *Guidance for generative AI in education and research*. UNESCO, 2023.

[21] *Ethics and Governance of Artificial Intelligence for Health : WHO Guidance*. World Health Organization, 2021.

[22] “Generative AI and the future of work The potential? Boundless. Deloitte AI Institute TM 2.” [Online]. Available: www.deloitte.com/us/AIInstitute

[23] M. Chui *et al.*, “The economic potential of generative AI The next productivity frontier The economic potential of generative AI: The next productivity frontier,” 2023.

[24] C. Liu, H. Wu, Y. Zhong, X. Zhang, Y. Wang, and W. Xie, “Intelligent Grimm -- Open-ended Visual Storytelling via Latent Diffusion Models,” Mar. 2024, [Online]. Available: http://arxiv.org/abs/2306.00973

[25] L. Yao, N. Peng, R. Weischedel, K. Knight, D. Zhao, and R. Yan, “Plan-And-Write: Towards Better Automatic Storytelling,” Feb. 2019, [Online]. Available: http://arxiv.org/abs/1811.05701

[26] C. Wu *et al.*, “Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation.” [Online]. Available: https://github.com/deepseek-ai/Janus

[27] Y. Jiao *et al.*, “UniToken: Harmonizing Multimodal Understanding and Generation through Unified Visual Encoding.”

[28] Y. Li *et al.*, “Uni-MoE: Scaling Unified Multimodal LLMs with Mixture of Experts,” May 2024, [Online]. Available: http://arxiv.org/abs/2405.11273

[29] A. Radford *et al.*, “Learning Transferable Visual Models From Natural Language Supervision,” Feb. 2021, [Online]. Available: http://arxiv.org/abs/2103.00020

[30] R. Girdhar *et al.*, “ImageBind: One Embedding Space To Bind Them All,” May 2023, [Online]. Available: http://arxiv.org/abs/2305.05665

[31] S. Li and H. Tang, “Multimodal Alignment and Fusion: A Survey,” Oct. 2025, [Online]. Available: http://arxiv.org/abs/2411.17040

[32] L. Zhang, A. Rao, and M. Agrawala, “Adding Conditional Control to Text-to-Image Diffusion Models,” *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3813–3824, Feb. 2023, doi: 10.1109/ICCV51070.2023.00355.

[33] M. Li *et al.*, “ControlNet++: Improving Conditional Controls with Efficient Consistency Feedback,” pp. 129–147, Apr. 2024, doi: 10.1007/978-3-031-72667-5_8.

[34] Y. Zhang, Y. Yuan, Y. Song, H. Wang, and J. Liu, “EasyControl: Adding Efficient and Flexible Control for Diffusion Transformer,” Mar. 2025, Accessed: Nov. 04, 2025. [Online]. Available: https://arxiv.org/pdf/2503.07027

[35] L. Zhang, A. Rao, and M. Agrawala, “Adding Conditional Control to Text-to-Image Diffusion Models,” *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3813–3824, Feb. 2023, doi: 10.1109/ICCV51070.2023.00355.

[36] C. Wang *et al.*, “Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers,” *IEEE Trans Audio Speech Lang Process*, vol. 33, pp. 705–718, Jan. 2023, doi: 10.1109/taslpro.2025.3530270.

[37] S. Chen *et al.*, “VALL-E 2: Neural Codec Language Models are Human Parity Zero-Shot Text to Speech Synthesizers,” Jun. 2024, Accessed: Nov. 04, 2025. [Online]. Available: https://arxiv.org/pdf/2406.05370

[38] Y. A. Li, X. Jiang, C. Han, and N. Mesgarani, “StyleTTS-ZS: Efficient High-Quality Zero-Shot Text-to-Speech Synthesis with Distilled Time-Varying Style Diffusion,” pp. 4725–4744, Sep. 2024, doi: 10.18653/v1/2025.naacl-long.242.

[39] T. Guan *et al.*, “HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models,” Mar. 2024, Accessed: Nov. 04, 2025. [Online]. Available: http://arxiv.org/abs/2310.14566

[40] G. Dagan, O. Loginova, and A. Batra, “CAST: Cross-modal Alignment Similarity Test for Vision Language Models,” pp. 1387–1402.

[41] Y. Chang, B. Cao, and L. Lin, “Monitoring Decoding: Mitigating Hallucination via Evaluating the Factuality of Partial Response during Generation,” pp. 14574–14587.

[42] P. Manakul, A. Liusie, and M. J. F. Gales, “SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models,” *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 9004–9017, Mar. 2023, doi: 10.18653/v1/2023.emnlp-main.557.

[43] Y. Gao *et al.*, “Retrieval-Augmented Generation for Large Language Models: A Survey,” Mar. 2024, [Online]. Available: http://arxiv.org/abs/2312.10997

[44] P. Manakul, A. Liusie, and M. J. F. Gales, “SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models,” *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 9004–9017, Mar. 2023, doi: 10.18653/v1/2023.emnlp-main.557.

[45] Z. Gou *et al.*, “CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing,” *12th International Conference on Learning Representations, ICLR 2024*, May 2023, Accessed: Nov. 04, 2025. [Online]. Available: https://arxiv.org/pdf/2305.11738

[46] H. Zhang *et al.*, “R-Tuning: Instructing Large Language Models to Say `I Don’t Know’,” *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024*, vol. 1, pp. 7106–7132, Nov. 2023, doi: 10.18653/v1/2024.naacl-long.394.

[47] Y. Tang and Y. Yang, “MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries”, Accessed: Nov. 04, 2025. [Online]. Available: https://github.com/yixuantt/MultiHop-RAG

[48] X. Chen, X. Hu, and N. Tang, “Review-Then-Refine: A Dynamic Framework for Multi-Hop Question Answering with Temporal Adaptability,” Dec. 2024, Accessed: Nov. 04, 2025. [Online]. Available: https://arxiv.org/pdf/2412.15101

[49] W. Xu, Z. Liang, K. Mei, H. Gao, J. Tan, and Y. Zhang, “A-Mem: Agentic Memory for LLM Agents”, Accessed: Nov. 04, 2025. [Online]. Available: https://github.com/WujiangXu/AgenticMemory

[50] Z. Zhang *et al.*, “A Survey on the Memory Mechanism of Large Language Model based Agents,” *ACM Trans Inf Syst*, vol. 43, no. 6, pp. 1–47, Apr. 2024, doi: 10.1145/3748302.

[51] M. Jia and Z. Sun, “A Survey of Multi-modal Emotion Recognition Based on Deep Learning,” *Highlights in Science, Engineering and Technology*, vol. 119, pp. 533–540, Dec. 2024, doi: 10.54097/37ZNCV36.

[52] M. P. A. Ramaswamy and S. Palaniswamy, “Multimodal emotion recognition: A comprehensive review, trends, and challenges,” *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 14, no. 6, p. e1563, Nov. 2024, doi: 10.1002/WIDM.1563;REQUESTEDJOURNAL:JOURNAL:19424795;PAGEGROUP:STRING:PUBLICATION.

[53] P. Bujnowski *et al.*, “SAMSEMO: New dataset for multilingual and multimodal emotion recognition,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2925–2929, 2024, doi: 10.21437/INTERSPEECH.2024-212.

[54] C. Y. Park *et al.*, “K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations,” *Sci Data*, vol. 7, no. 1, May 2020, doi: 10.1038/s41597-020-00630-y.

[55] H. Sun *et al.*, “EmotionTalk: An Interactive Chinese Multimodal Emotion Dataset With Rich Annotations,” May 2025, Accessed: Nov. 04, 2025. [Online]. Available: https://arxiv.org/pdf/2505.23018

[56] S. Pawar *et al.*, “Survey of Cultural Awareness in Language Models: Text and Beyond,” Oct. 2024, Accessed: Nov. 04, 2025. [Online]. Available: https://arxiv.org/pdf/2411.00860v1

[57] A. Gulati, M. D’incà, N. Sebe, B. L. Fondazione, B. Kessler, and N. Oliver, “Beauty and the Bias: Exploring the Impact of Attractiveness on Multimodal Large Language Models,” Apr. 2025, Accessed: Nov. 04, 2025. [Online]. Available: https://arxiv.org/pdf/2504.16104

[58] S. Nayak *et al.*, “Benchmarking Vision Language Models for Cultural Understanding,” *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 5769–5790, Jul. 2024, doi: 10.18653/v1/2024.emnlp-main.329.

[59] J. S. Kim *et al.*, “When Tom Eats Kimchi: Evaluating Cultural Bias of Multimodal Large Language Models in Cultural Mixture Contexts,” Mar. 2025, Accessed: Nov. 04, 2025. [Online]. Available: https://arxiv.org/pdf/2503.16826

[60] Y. Wang and M. Kreminski, “Can LLMs Generate Good Stories? Insights and Challenges from a Narrative Planning Perspective,” Jun. 2025, Accessed: Nov. 04, 2025. [Online]. Available: https://arxiv.org/pdf/2506.10161v1

[61] S. Yang *et al.*, “SEED-Story: Multimodal Long Story Generation with Large Language Model,” Oct. 2024, [Online]. Available: http://arxiv.org/abs/2407.08683

[62] T. Rahman, H.-Y. Lee, J. Ren, S. Tulyakov, S. Mahajan, and L. Sigal, “Make-A-Story: Visual Memory Conditioned Consistent Story Generation,” May 2023, [Online]. Available: http://arxiv.org/abs/2211.13319

[63] Z. Guo, F. Zhang, K. Jia, and T. Jin, “LLM-I: LLMs are Naturally Interleaved Multimodal Creators,” Sep. 2025, [Online]. Available: http://arxiv.org/abs/2509.13642

[64] J. Wu, C. Tang, J. Wang, Y. Zeng, X. Li, and Y. Tong, “DiffSensei: Bridging Multi-Modal LLMs and Diffusion Models for Customized Manga Generation,” Mar. 2025, [Online]. Available: http://arxiv.org/abs/2412.07589

[65] Y. Zhang *et al.*, “STICKERCONV: Generating Multimodal Empathetic Responses from Scratch,” Feb. 2024, doi: 10.48448/6s4n-d973.

[66] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, “HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face,” Mar. 2023, [Online]. Available: http://arxiv.org/abs/2303.17580

[67] X. Chen *et al.*, “EmpathyAgent: Can Embodied Agents Conduct Empathetic Actions?,” Mar. 2025, [Online]. Available: http://arxiv.org/abs/2503.16545