Capítulo 5

Agentes de Inteligência Artificial Aplicados à Saúde: Fundamentos, Arquiteturas e Implementação Prática

James D. Sousa, Frank Cesár Lopes Véras

Resumo

A Inteligência Artificial vêm transformando sistemas de saúde ao possibilitar automação, suporte à decisão clínica, monitoramento de pacientes e intervenções personalizadas. Nesse contexto, os agentes de IA emergem como uma evolução natural dos sistemas inteligentes, oferecendo maior autonomia, adaptabilidade e capacidade de interação. Este minicurso apresenta fundamentos teóricos e implementação prática de agentes de IA aplicados à saúde. Serão abordados modelos de linguagem de larga escala (LLMs), arquiteturas de agentes autônomos, geração aumentada por recuperação (RAG), sistemas de apoio à decisão clınica, integração de dados clınicos e aspectos éticos. Os participantes desenvolverão um protótipo funcional de agente de IA aplicado a cenários simulados da saúde. O objetivo é integrar teoria e prática no desenvolvimento de soluções digitais para saúde.

Abstract

Artificial Intelligence has been transforming healthcare systems by enabling automation, clinical decision support, patient monitoring, and personalized interventions. In this context, AI agents emerge as a natural evolution of intelligent systems, offering greater autonomy, adaptability, and interaction capabilities. This minicourse presents the theoretical foundations and practical implementation of AI agents applied to healthcare. Topics include large language models (LLMs), autonomous agent architectures, retrieval-augmented generation (RAG), clinical decision support systems, integration of clinical data, and ethical considerations. Participants will develop a functional prototype of an AI agent applied to simulated healthcare scenarios. The objective is to integrate theory and practice in the development of digital healthcare solutions.

5.1. Introdução

A Inteligência Artificial (IA) pode ser compreendida como o campo da ciência da computação dedicado ao desenvolvimento de sistemas capazes de realizar tarefas que, até recentemente, exigiam exclusivamente a cognição humana como reconhecer padrões, tomar decisões, compreender linguagem e aprender com experiências passadas. Desde suas origens teóricas nas décadas de 1950 e 1960, a IA evoluiu de experimentos acadêmicos isolados para uma das forças mais transformadoras da economia global, presente em setores que vão da indústria ao entretenimento, das finanças à logística.

A incorporação da IA no setor da saúde representa uma das transformações mais significativas da era digital. Com o avanço das tecnologias computacionais e o aumento da disponibilidade de dados clínicos, tornou-se possível desenvolver sistemas capazes de auxiliar profissionais da saúde em tarefas complexas, desde diagnósticos até a gestão hospitalar (ESTEVA et al., 2019; TOPOL, 2019). Segundo a pesquisa TIC Saúde, 17% dos médicos brasileiros utilizam alguma forma de IA em sua prática clínica, número que chega a 20% nas instituições privadas (Núcleo de Informação e Coordenação do Ponto BR (NIC.br), 2024).

Levantamento da Associação Nacional de Hospitais Privados (Anahp) aponta que 62,5% dos hospitais privados já integram IA em suas operações (Associação Nacional de Hospitais Privados (Anahp), 2025). Globalmente, o mercado de healthtechs na América Latina registrou crescimento de 37,6% nos investimentos em 2024, atingindo US$ 253,7 milhões reflexo da confiança crescente da indústria no potencial transformador dessas tecnologias (Distrito; Associação Brasileira de Startups de Saúde e HealthTechs (ABSS), 2025).

5.1.1. Definição de Agente Inteligente

Um agente inteligente pode ser definido como uma entidade computacional capaz de perceber o ambiente em que está inserida, processar essas informações e agir de forma autônoma para atingir determinados objetivos. Essa definição envolve três componentes fundamentais: percepção, decisão e ação (RUSSELL; NORVIG; INTELLIGENCE, 1995; WOOLDRIDGE, 2009). No contexto da saúde, a percepção pode envolver a coleta de dados clínicos, como sinais vitais, exames laboratoriais e histórico do paciente. A etapa de decisão está relacionada à análise desses dados, frequentemente utilizando algoritmos de aprendizado de máquina ou modelos probabilísticos.

Por fim, a ação pode se manifestar na forma de recomendações clínicas, alertas ou até mesmo intervenções automatizadas (REZENDE, 2003; GOODFELLOW et al., 2016). Diferentemente de sistemas tradicionais, agentes inteligentes podem operar em ambientes dinâmicos e incertos, ajustando seu comportamento com base em novas informações. Essa característica é particularmente relevante na área da saúde, onde decisões precisam ser tomadas com base em dados incompletos e em constante atualização.

O Hospital Israelita Albert Einstein[1], por exemplo, desenvolveu o *HStory*, uma plataforma que utiliza um agente de IA generativa para consolidar dados de pacientes distribuídos em diferentes sistemas em um único painel centralizado, sintetizando automa-

[1]https://www.einstein.br/n

ticamente o histórico clínico e devolvendo ao médico mais tempo para o cuidado direto. Trata-se de um agente que percebe dados dispersos no ambiente hospitalar exames, internações, prescrições e os transforma em informação acionável, sem depender de intervenção humana em cada etapa do processo.

5.1.2. Evolução Histórica da IA na Saúde

A aplicação da IA na saúde não é recente. Desde a década de 1970 sistemas especialistas já eram utilizados para auxiliar no diagnóstico médico. Um exemplo clássico é o MYCIN[2], desenvoolvido pela universidade de stanford, que utilizava regras baseadas no conhecimento de especialistas para identificar infecções bacterianas e recomendar tratamentos (SHORTLIFFE, 2012). O MYCIN demonstrou que era possível codificar o raciocínio clínico em regras lógicas, mas também revelou as limitações dessa abordagem sistemas rígidos que não aprendiam com novos dados e dependiam de atualização manual constante.

Com o avanço do poder computacional e o surgimento de grandes bases de dados, a IA evoluiu para incorporar técnicas de aprendizado de máquina, permitindo que sistemas aprendessem padrões diretamente dos dados. Essa mudança marcou a transição de sistemas baseados em regras para sistemas baseados em dados. Redes neurais artificiais, Support Vector Machines e métodos de ensemble passaram a dominar a pesquisa, alcançando resultados expressivos em tarefas como classificação de exames de imagem e predição de risco clínico.

Mais recentemente, o desenvolvimento de modelos de linguagem avançados trouxe uma nova dimensão para a IA na saúde. Esses modelos são capazes de compreender linguagem natural, interpretar textos médicos e interagir com usuários de forma mais intuitiva. Como resultado, surgiram agentes mais sofisticados, capazes de integrar múltiplas fontes de informação e oferecer suporte mais abrangente. Além disso, a crescente digitalização dos serviços de saúde, incluindo prontuários eletrônicos e dispositivos vestíveis, ampliou significativamente o volume de dados disponíveis. Isso possibilitou o desenvolvimento de agentes que não apenas reagem a eventos, mas também antecipam situações, contribuindo para uma abordagem mais preventiva no contexto clínico.

Portanto, a trajetória da IA na saúde pode ser dividida em grandes fases. Nos anos 1970, sistemas especialistas como MYCIN e INTERNIST-1 aplicavam lógica baseada em regras(SHORTLIFFE, 2012; MILLER; JR; MYERS, 1985). Nas décadas de 1980 e 1990, redes neurais artificiais começaram a ser aplicadas à análise de imagens médicas e sinais biomédicos(LIPPMANN, 1988). Nos anos 2000, o aprendizado de máquina e a mineração de dados em grandes bases clínicas dominaram a pesquisa(HAN; KAMBER; PEI, 2011). Nos anos 2010, o deep learning revolucionou o diagnóstico por imagem e o IBM Watson Health trouxe a IA para o debate clínico mainstream(LECUN; BENGIO; HINTON, 2015). Em 2017, a arquitetura Transformer mudou definitivamente o paradigma dos modelos de (VASWANI et al., 2017). Nos anos 2020, LLMs, agentes autônomos, RAG e o MCP consolidam uma nova geração de assistentes clínicos (LEWIS et al., 2020; ANTHROPIC, 2024).

[2]<https://www.historytools.org/inventions/mycin-expert-system-the-first-ai-medical-diagnosis>

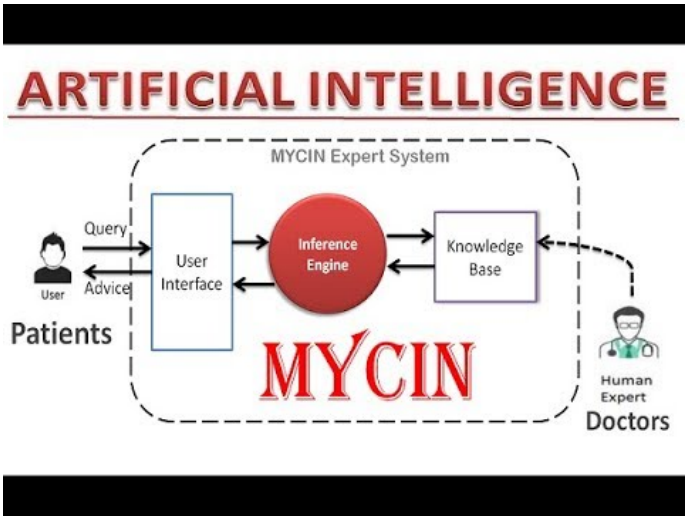

Figura 5.1. Aquitetura do MYCIN

5.1.3. Chatbot vs. Agente Autônomo

Embora os termos *chatbot* e agente autônomo sejam frequentemente utilizados de forma intercambiável no discurso popular, eles descrevem arquiteturas e capacidades fundamentalmente distintas. Compreender essa diferença é essencial para avaliar o potencial e as limitações de cada abordagem no contexto da saúde digital (Figura 5.9). Os chatbots tradicionais são sistemas de diálogo baseados em regras predefinidas ou em modelos estatísticos simples. Eles operam dentro de fluxos conversacionais estruturados, respondendo a intenções detectadas em mensagens dos usuários com respostas pré-cadastradas ou geradas a partir de *templates*. No contexto da saúde, chatbots desse tipo têm sido utilizados para triagem sintomática, agendamento de consultas e fornecimento de informações básicas de saúde pública.

Agentes autônomos, por sua vez, operam em um nível de sofisticação radicalmente superior. Eles são capazes de planejar sequências de ações, utilizar ferramentas externas, manter memória de interações anteriores e adaptar seu comportamento com base no contexto. Em vez de seguir fluxos predefinidos, um agente autônomo raciocina sobre o problema, decide quais recursos consultar e executa ações complexas com mínima intervenção humana. Na saúde, essa distinção tem implicações diretas enquanto um chatbot pode informar o horário de funcionamento de uma clínica, um agente autônomo pode analisar o histórico clínico de um paciente, consultar diretrizes atualizadas, identificar contraindicações e sugerir um plano de cuidado fundamentado em evidências.

Essa diferença se manifesta em diversas dimensões práticas. Em termos de arquitetura, chatbots seguem fluxos predefinidos com regras fixas, enquanto agentes operam por raciocínio dinâmico e planejamento adaptativo. Quanto à memória, chatbots geralmente não retêm informações entre sessões, ao passo que agentes mantêm memória de curto e longo prazo. No que diz respeito ao uso de ferramentas externas, chatbots raramente se integram a sistemas externos, enquanto agentes acessam APIs, bases de dados e prontuários de forma autônoma. Na prática clínica, essa distinção se traduz em capacidades radicalmente distintas um chatbot realiza triagem básica e responde a perguntas frequentes

um agente como Ada Health, DAX Copilot ou HStory conduz raciocínio diagnóstico, personaliza recomendações e documenta automaticamente o atendimento.

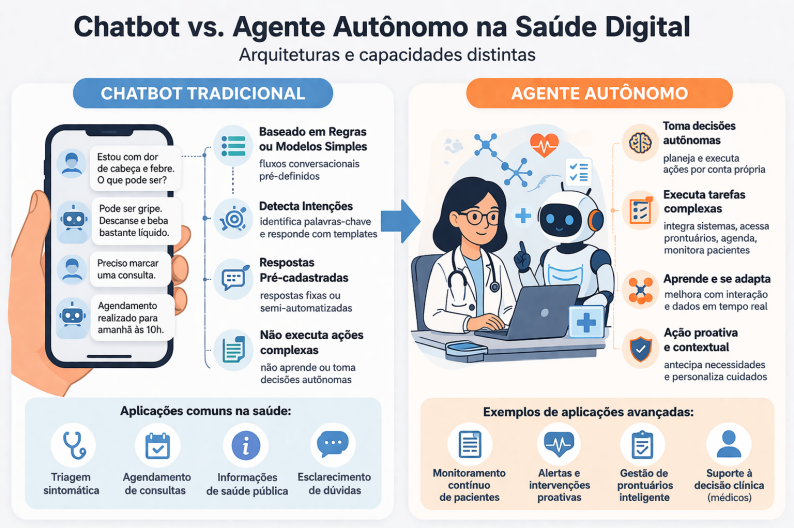

Figura 5.2. Chatbot x Agente

5.2. Como os Agentes de IA Funcionam

Compreender o funcionamento interno de um agente de IA é fundamental para avaliar seu potencial e suas limitações. Diferentemente de sistemas convencionais, que executam instruções predefinidas em sequências fixas, os agentes inteligentes modernos operam por meio de arquiteturas dinâmicas, capazes de perceber o ambiente, raciocinar sobre ele e agir de forma autônoma e adaptativa. Esta seção percorre os três pilares que sustentam esse funcionamento a arquitetura clássica de percepção – planejamento – ação, os modelos de linguagem de larga escala que viabilizam o raciocínio em linguagem natural, e as técnicas de recuperação de informação que permitem ao agente acessar e integrar conhecimento clínico especializado em tempo real.

5.2.1. Arquitetura Percepção – Planejamento – Ação

A compreensão do funcionamento interno de agentes inteligentes é fundamental para o desenvolvimento de sistemas autônomos robustos e confiáveis. Entre os diversos modelos propostos na literatura, a arquitetura baseada no ciclo percepção – planejamento – ação destaca-se como uma das mais clássicas e amplamente utilizadas (MÜLLER, 1999). Esse modelo descreve o comportamento de um agente como um processo contínuo de interação com o ambiente, no qual ele percebe informações, toma decisões e executa ações com o objetivo de atingir determinadas metas.

Esse paradigma tem origem em estudos da área de Inteligência Artificial e sistemas autônomos, sendo discutido em obras clássicas como as de Stuart Russell e Peter Norvig (RUSSELL; NORVIG; INTELLIGENCE, 1995), que estruturaram formalmente o conceito de agentes inteligentes como entidades que percebem o ambiente por meio de sensores e agem sobre ele por meio de atuadores.

5.2.2. Percepção

A primeira etapa do ciclo corresponde à percepção, que envolve a coleta e interpretação de dados provenientes do ambiente. Esses dados podem assumir diversas formas, como texto, imagens, sinais biomédicos, registros clínicos ou interações do usuário. Em sistemas computacionais, essa etapa é geralmente implementada por meio de sensores lógicos (APIs, entradas de usuário, arquivos digitais) ou sensores físicos (dispositivos IoT, equipamentos médicos, etc.).

No contexto da saúde digital, por exemplo, a percepção pode envolver a leitura de prontuários eletrônicos, resultados de exames laboratoriais ou sinais vitais monitorados em tempo real. O agente precisa não apenas coletar esses dados, mas também interpretá-los corretamente, o que frequentemente exige técnicas de processamento de linguagem natural, visão computacional ou análise de séries temporais. Além disso, a percepção pode incluir etapas de pré-processamento, como limpeza de dados, normalização e extração de características relevantes. Essa fase é crítica, pois a qualidade das decisões do agente depende diretamente da qualidade das informações percebidas.

5.2.3. Planejamento

Após perceber o ambiente, o agente entra na fase de planejamento, na qual decide quais ações devem ser executadas para atingir seus objetivos. Essa etapa envolve raciocínio, tomada de decisão e, em muitos casos, previsão de cenários futuros. O planejamento pode variar em complexidade, desde regras simples (por exemplo, sistemas baseados em "se - então"), até algoritmos mais sofisticados, como busca heurística, algoritmos de otimização e aprendizado de máquina. Em arquiteturas mais avançadas, o agente pode manter um modelo interno do ambiente, permitindo simular diferentes estratégias antes de agir.

No contexto de agentes modernos baseados em IA, o planejamento também pode envolver o uso de modelos probabilísticos e técnicas de inferência, permitindo lidar com incertezas e informações incompletas. Em aplicações na saúde, por exemplo, o agente pode avaliar diferentes hipóteses diagnósticas com base nos dados disponíveis e selecionar a mais provável. Outro aspecto importante do planejamento é a definição de objetivos e restrições. Em ambientes críticos, como sistemas clínicos, as decisões precisam respeitar diretrizes médicas, protocolos de segurança e aspectos éticos, o que torna o planejamento ainda mais desafiador.

5.2.4. Ação

A terceira etapa do ciclo é a ação, na qual o agente executa as decisões tomadas durante o planejamento. As ações podem variar desde a geração de uma resposta textual até a execução de comandos em sistemas externos, como atualização de registros, envio de alertas ou controle de dispositivos. Em sistemas baseados em software, as ações geralmente são realizadas por meio de APIs, chamadas de sistema ou interfaces com outros serviços. Já em sistemas físicos, como robôs ou dispositivos médicos, as ações podem envolver atuadores que interagem diretamente com o ambiente.

Na área da saúde, exemplos de ações incluem a recomendação de tratamentos, o envio de alertas para profissionais de saúde ou a priorização de atendimentos com base na gravidade dos pacientes. É importante destacar que, nesses contextos, a execução

de ações deve ser cuidadosamente controlada, muitas vezes exigindo validação humana (*human-in-the-loop*) para garantir segurança e confiabilidade.

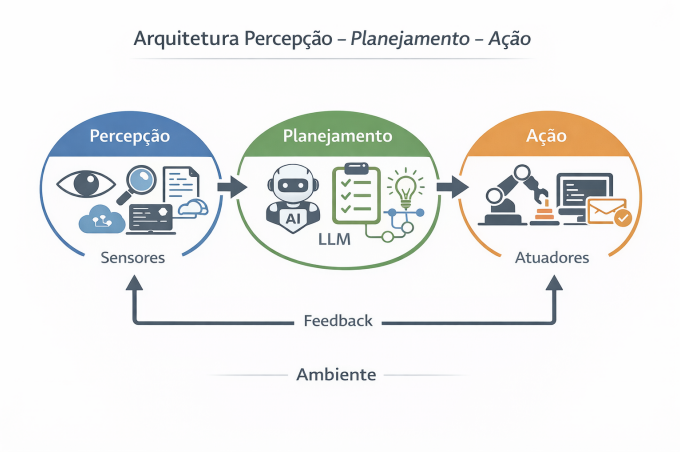

Figura 5.3. Arquitetura PPA

5.3. Memória dos Agentes

Uma dimensão frequentemente subestimada na arquitetura de agentes é a memória. Agentes de IA modernos operam com dois tipos principais de memória, cada um com papéis distintos no suporte à decisão clínica.

A memória de curto prazo corresponde ao contexto da sessão atual, incluindo o histórico de mensagens trocadas, os documentos já consultados e as decisões tomadas durante aquela interação específica. Trata-se de uma memória volátil, pois as informações normalmente são descartadas ao término da sessão. Em um agente clínico, esse tipo de memória permite que o sistema lembre, por exemplo, que o paciente mencionou alergia à penicilina no início da conversa ao formular uma recomendação terapêutica vários turnos depois. A capacidade da janela de contexto, geralmente medida em *tokens*, determina o quanto o agente consegue manter simultaneamente em processamento.

Já a memória de longo prazo refere-se a informações persistidas entre sessões, normalmente armazenadas em bancos de dados vetoriais ou relacionais. Esse mecanismo permite que o agente recupere interações anteriores com o mesmo paciente, acumule preferências institucionais e mantenha um histórico clínico continuamente consultável. Além disso, um terceiro tipo emergente vem ganhando relevância, a chamada memória episódica. Esse modelo permite ao agente recuperar interações semanticamente relevantes do passado, mesmo que não sejam as mais recentes. Em ambientes hospitalares, essa capacidade é particularmente importante para garantir continuidade do cuidado, sobretudo em pacientes crônicos que possuem múltiplas internações e tratamentos distribuídos ao longo do tempo.

5.4. Modelos de Linguagem de Larga Escala (LLMs)

Os Modelos de Linguagem de Larga Escala (*Large Language Models* – LLMs) representam a principal tecnologia responsável pela nova geração de agentes inteligentes. Esses

sistemas utilizam técnicas de aprendizado profundo treinadas em enormes volumes de texto, permitindo compreender e gerar linguagem natural com elevado grau de sofisticação. Entre os exemplos mais conhecidos estão modelos desenvolvidos por organizações como a OpenAI[3] e o Google AI[4], que demonstraram avanços significativos na interação entre humanos e máquinas. Para compreender o papel desses modelos em agentes aplicados à saúde, é necessário entender sua arquitetura subjacente e o processo pelo qual adquirem conhecimento.

A arquitetura *Transformer*, proposta por Vaswani et al. (VASWANI et al., 2017), constitui a base dos modelos de linguagem modernos. Seu principal diferencial é o mecanismo de atenção, que permite ao modelo identificar quais partes de um texto são mais relevantes em determinado contexto. Esse mecanismo possibilita capturar relações semânticas entre palavras mesmo quando estão distantes umas das outras, superando limitações observadas em arquiteturas anteriores, como redes neurais recorrentes. Na prática, essa característica é particularmente útil em aplicações médicas. Ao analisar um prontuário clínico extenso, por exemplo, o modelo consegue relacionar informações mencionadas anteriormente como uma alergia à penicilina com decisões terapêuticas tomadas posteriormente durante a consulta.

O treinamento dos LLMs geralmente ocorre em duas etapas principais. Na primeira, denominada pré-treinamento, o modelo aprende padrões linguísticos a partir da leitura de grandes volumes de textos, incluindo livros, artigos científicos e conteúdos disponíveis na internet. Em seguida, ocorre o ajuste fino (*fine-tuning*) ou a aplicação de técnicas como o RLHF (*Reinforcement Learning from Human Feedback*), nas quais o modelo é refinado para responder de maneira mais útil, segura e alinhada às instruções humanas, inclusive em domínios específicos como a medicina.

O uso de LLMs na área médica avançou rapidamente nos últimos anos. Um exemplo relevante é o Med-PaLM 2[5], desenvolvido pelo Google, que alcançou desempenho comparável ao de médicos em avaliações relacionadas ao *United States Medical Licensing Examination* (USMLE). Outro exemplo importante é o BioMedLM[6], desenvolvido para aplicações biomédicas específicas, além do ClinicalBERT[7], treinado com prontuários clínicos e literatura médica. Esses modelos especializados apresentam desempenho significativo em tarefas como extração de informações clínicas, apoio ao diagnóstico e triagem de pacientes.

A principal vantagem desses modelos especializados é a capacidade de compreender terminologias médicas, abreviações clínicas e padrões comuns encontrados em prontuários e exames. Dessa forma, conseguem operar com maior precisão em ambientes clínicos quando comparados a modelos generalistas. Entretanto, ainda existem limitações importantes. Como dependem dos dados utilizados durante o treinamento, esses sistemas podem não possuir conhecimento atualizado sobre eventos recentes ou novas diretrizes médicas. Além disso, podem produzir respostas incorretas com alto grau aparente de confi-

[3]<https://openai.com>
[4]<https://ai.google>
[5]<https://sites.research.google/med-palm/>
[6]<https://crfm.stanford.edu/2022/12/15/biomedlm.html>
[7]<https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT>

ança, fenômeno conhecido como "alucinação". Por esse motivo, recomenda-se que esses sistemas sejam utilizados em conjunto com fontes confiáveis de informação e mecanismos adicionais de verificação.

Nos agentes inteligentes, os LLMs funcionam como o núcleo de raciocínio do sistema. Eles recebem simultaneamente informações provenientes do ambiente, histórico de conversas, resultados de buscas externas e instruções operacionais, produzindo respostas em linguagem natural ou ações estruturadas, como chamadas de ferramentas externas. Essa capacidade de combinar raciocínio contextual com uso de ferramentas torna os agentes particularmente úteis em cenários complexos, como o ambiente hospitalar.

Um exemplo prático dessa aplicação é o Nabla Copilot[8]. Durante consultas médicas, o sistema processa a transcrição da conversa, consulta informações do prontuário do paciente e utiliza diretrizes médicas para gerar automaticamente um rascunho estruturado da consulta em tempo real. Ao final do processo, o profissional de saúde revisa e valida as informações antes do registro definitivo, mantendo a supervisão humana sobre a decisão clínica. Essa abordagem evidencia como a colaboração entre profissionais e sistemas de IA pode aumentar a eficiência operacional sem comprometer a responsabilidade médica e a qualidade do atendimento.

5.5. Retrieval-Augmented Generation (RAG)

Uma das principais limitações dos LLMs, especialmente em áreas críticas como a saúde, é que eles não têm acesso a informações atualizadas automaticamente. Eles aprendem com dados até um certo momento e, depois disso, não sabem de novas diretrizes médicas, estudos recentes, novos medicamentos ou até informações específicas de um paciente que estão em sistemas hospitalares. Para resolver esse problema, surgiu uma abordagem chamada *Retrieval-Augmented Generation (RAG)*. De forma simples, o RAG combina duas coisas: a capacidade do modelo de gerar respostas com a habilidade de buscar informações atualizadas em fontes externas. Em vez de depender só do que aprendeu no passado, o sistema procura dados relevantes em tempo real como documentos, bases médicas ou registros e usa essas informações para construir respostas mais corretas e atualizadas.

5.5.1. Como o RAG funciona: o pipeline de recuperação

O funcionamento do Retrieval-Augmented Generation (RAG) pode ser entendido de forma simples como um processo em etapas. Primeiro, os documentos importantes como diretrizes médicas, artigos científicos e protocolos são organizados e transformados em uma espécie de "linguagem numérica" que representa o significado dos textos. Isso permite que o sistema entenda melhor o conteúdo, e tudo fica armazenado em um banco de dados preparado para esse tipo de busca. Depois, quando alguém faz uma pergunta, o sistema transforma essa pergunta nesse mesmo formato e procura, dentro da base, os trechos mais parecidos em termos de significado. Ou seja, ele não busca só por palavras iguais, mas por conteúdo relevante. Em seguida, esses trechos encontrados são enviados junto com a pergunta para o modelo. Assim, o LLM não precisa depender apenas do que aprendeu no passado ele passa a usar informações atualizadas naquele momento. Por fim, o modelo gera uma resposta com base nesses documentos. Em muitos casos, ele pode até indicar

[8]<https://www.nabla.com/copilot>

de onde tirou as informações, o que é muito importante em áreas como a saúde, onde é essencial saber a origem dos dados e garantir mais segurança nas decisões.

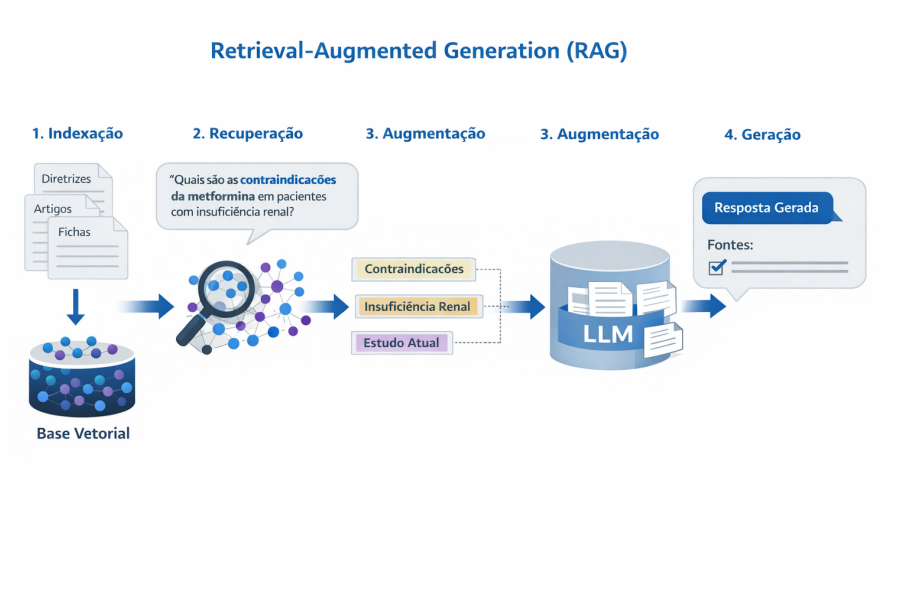

Figura 5.4. Aquitetura Rag

5.5.2. Tipos de Recuperação e Recuperação Híbrida

Existem dois paradigmas principais de recuperação de informação em sistemas baseados em RAG. O primeiro é a busca densa (*dense retrieval*), fundamentada em embeddings vetoriais e similaridade semântica. Nesse modelo, documentos e consultas são representados como vetores em um espaço multidimensional, permitindo recuperar conteúdos semanticamente relacionados mesmo quando não há correspondência literal entre os termos utilizados. O segundo paradigma é a busca esparsa (*sparse retrieval*), baseada em métodos tradicionais como TF-IDF e BM25, nos quais a recuperação ocorre principalmente por correspondência lexical entre palavras presentes na consulta e nos documentos.

Cada abordagem apresenta vantagens e limitações específicas. A busca densa tende a capturar melhor a intenção semântica da consulta, sendo particularmente útil quando diferentes termos podem representar conceitos semelhantes. Já a busca esparsa oferece elevada precisão em cenários nos quais termos técnicos exatos são fundamentais, como nomes de medicamentos, códigos CID, resultados laboratoriais ou nomenclaturas de procedimentos médicos.

Para superar as limitações individuais de cada abordagem, muitos sistemas modernos utilizam recuperação híbrida (*hybrid retrieval*), combinando simultaneamente busca densa e esparsa. Após a etapa inicial de recuperação, técnicas de *re-ranking* podem ser aplicadas para reorganizar os documentos encontrados conforme critérios adicionais de relevância semântica e contextual antes que o conteúdo seja encaminhado ao modelo de linguagem. Essa estratégia melhora significativamente a qualidade das respostas geradas, especialmente em ambientes clínicos, nos quais precisão terminológica e compreensão contextual são igualmente importantes.

5.5.3. Vantagens do RAG no contexto clínico

O uso do Retrieval-Augmented Generation na área da saúde traz varios benefícios que vão além da simples atualização do conhecimento. Em primeiro lugar, essa técnica reduz significativamente o risco de alucinações: quando o modelo é explicitamente instruído a responder com base apenas nos documentos recuperados, a tendência de gerar informações não fundamentadas diminui de forma substancial.

Em segundo lugar, o RAG permite a personalização do agente para contextos institucionais específicos: um hospital pode indexar seus próprios protocolos, formulários e diretrizes locais, garantindo que o agente opere com base nas práticas da instituição e não em recomendações genéricas.

Em terceiro lugar, o RAG viabiliza a integração de dados do paciente em tempo real. Ao indexar os registros clínicos de um paciente específico incluindo histórico de exames, diagnósticos anteriores, alergias e medicamentos em uso, o agente pode recuperar essas informações contextualmente relevantes e incorporá-las ao raciocínio clínico. Esse modelo é o que sustenta, por exemplo, plataformas como o Suki AI, assistente de voz para médicos amplamente adotado nos Estados Unidos, que recupera informações do prontuário do paciente para contextualizar automaticamente as notas clínicas geradas durante a consulta.

5.6. Uso de Ferramentas Externas e o Model Context Protocol (MCP)

Agentes de IA modernos não operam de forma isolada. Para executar tarefas complexas, precisam interagir continuamente com sistemas externos, realizando operações como consultar bases de dados clínicas, acessar APIs laboratoriais, verificar disponibilidade de leitos, agendar exames e recuperar informações de prontuários eletrônicos. Essa capacidade de utilizar ferramentas externas frequentemente denominada *tool use* ou *function calling* constitui uma das principais características que diferenciam agentes autônomos de simples modelos geradores de texto.

Historicamente, integrações entre sistemas eram desenvolvidas de maneira *ad hoc*, utilizando código customizado, altamente acoplado e de difícil manutenção. Em ambientes hospitalares, essa limitação torna-se ainda mais evidente devido à coexistência de múltiplos sistemas heterogêneos, como HIS (*Hospital Information Systems*), LIS (*Laboratory Information Systems*), RIS (*Radiology Information Systems*) e plataformas de farmácia hospitalar. Além disso, esses sistemas frequentemente utilizam padrões distintos de comunicação, incluindo HL7 v2[9], FHIR[10] e protocolos proprietários.

Essa fragmentação tecnológica representa um obstáculo significativo para a adoção de agentes inteligentes em larga escala na saúde, pois dificulta interoperabilidade, padronização e manutenção das integrações. Consequentemente, grande parte do esforço de desenvolvimento desses agentes acaba sendo direcionada não apenas à inteligência do modelo, mas também à criação de mecanismos robustos de comunicação entre diferentes sistemas clínicos e administrativos.

[9] <https://www.hl7.org/implement/standards/product_brief.cfm?product_id=185>
[10] <https://www.hl7.org/fhir/>

5.6.1. O Que é o MCP

O *Model Context Protocol* (MCP) é um protocolo aberto lançado pela Anthropic em 2024 que padroniza a forma como LLMs se comunicam com ferramentas e fontes de dados externas. Frequentemente descrito como um "USB para agentes de IA", o MCP define uma interface comum: qualquer ferramenta ou sistema que implemente o protocolo pode ser conectado a qualquer agente compatível, sem necessidade de integrações customizadas. Desde seu lançamento, o protocolo foi adotado por dezenas de plataformas, incluindo IDEs de desenvolvimento, ferramentas de produtividade e, progressivamente, sistemas de informação em saúde.

A arquitetura do MCP é composta por três elementos principais. O *MCP Host* corresponde à aplicação do agente, responsável por orquestrar as interações. O *MCP Client* gerencia as conexões entre o host e os servidores. Já o *MCP Server* expõe ferramentas e fontes de dados específicas, como servidores responsáveis pelo acesso a prontuários eletrônicos, consulta a bases farmacológicas ou verificação da disponibilidade de leitos hospitalares.

5.6.2. Relevância Clínica do MCP

No contexto hospitalar, o MCP possui implicações práticas imediatas. Um agente de apoio à decisão clínica pode, por meio de servidores MCP, consultar resultados de exames em tempo real, verificar interações medicamentosas em bases farmacológicas como o DrugBank[11], checar disponibilidade de especialistas para teleconsultas e registrar evoluções diretamente no prontuário eletrônico. Tudo isso ocorre utilizando um protocolo padronizado de comunicação, reduzindo significativamente a necessidade de integrações proprietárias específicas para cada sistema hospitalar.

Além dos ganhos operacionais, o MCP oferece vantagens relevantes sob a perspectiva regulatória. Ao estabelecer uma camada explícita e padronizada de comunicação entre agentes inteligentes e sistemas externos, o protocolo facilita auditoria, rastreabilidade e monitoramento das ações executadas pelo agente. Em um ambiente altamente regulado como o setor de saúde, no qual rastreabilidade representa simultaneamente uma exigência ética e legal especialmente diante da Lei Geral de Proteção de Dados (LGPD)[12] e das regulamentações da ANVISA[13] essa característica torna-se particularmente valiosa.

Apesar de seu potencial, o MCP ainda é considerado um protocolo relativamente recente. O ecossistema de servidores especializados para sistemas de informação em saúde permanece em rápida evolução, com iniciativas voltadas à integração de prontuários eletrônicos, plataformas laboratoriais, sistemas de imagem médica e serviços de suporte clínico baseados em IA. Além disso, desafios importantes ainda precisam ser superados para que o protocolo alcance ampla adoção em ambientes hospitalares reais. Entre esses desafios destacam-se questões relacionadas à padronização semântica dos dados clínicos, interoperabilidade entre sistemas legados, controle de permissões de acesso e garantia de privacidade das informações sensíveis dos pacientes. Em muitos hospitais, parte significativa da infraestrutura ainda depende de softwares antigos e altamente personalizados, o que

[11] <https://go.drugbank.com>

[12] <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>

[13] <https://www.gov.br/anvisa/pt-br>

dificulta a adoção de arquiteturas modernas baseadas em agentes inteligentes e protocolos padronizados.

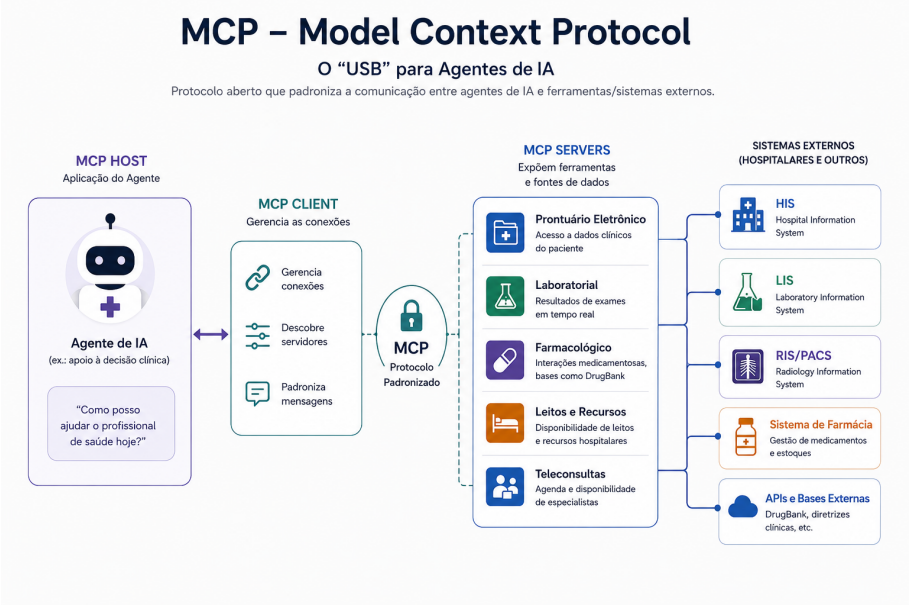

Figura 5.5. MCP

5.7. Sistemas Multi-Agentes

Em cenários clínicos de elevada complexidade, um único agente inteligente pode não ser suficiente para lidar adequadamente com todas as demandas operacionais e cognitivas do ambiente hospitalar. Nesse contexto, surgem os sistemas multiagentes (*Multi-Agent Systems* – MAS), nos quais múltiplos agentes especializados distribuem responsabilidades e colaboram para resolver problemas de maneira coordenada. Cada agente possui competências específicas e delimitadas, comunicando-se com os demais por meio de protocolos definidos e mecanismos estruturados de troca de informações.

Um exemplo aplicado à saúde pode ser observado em ambientes de pronto-socorro. Um agente de triagem seria responsável por classificar pacientes conforme critérios de gravidade e encaminhar os casos para agentes especializados. Pacientes com dor torácica poderiam ser direcionados a um agente cardiológico, enquanto quadros com déficit neurológico focal seriam encaminhados a um agente neurológico. Paralelamente, um agente de gestão hospitalar poderia monitorar disponibilidade de leitos e coordenar processos de internação quando necessário. Embora cada agente opere com relativa autonomia dentro de sua especialidade, o sistema global funciona de maneira integrada e colaborativa.

A coordenação desses ambientes multiagentes é viabilizada por *frameworks* especializados, como LangGraph[14], AutoGen[15] e CrewAI[16]. Essas plataformas oferecem infraestrutura para orquestração de agentes, definição de grafos de dependência, gerenciamento de fluxo de informações e controle das interações entre diferentes componentes do

[14]<https://www.langchain.com/langgraph>
[15]<https://microsoft.github.io/autogen/>
[16]<https://www.crewai.com>

sistema. Em aplicações médicas, essa abordagem permite distribuir tarefas complexas entre múltiplos agentes especializados, aumentando escalabilidade, modularidade e capacidade de tomada de decisão em ambientes clínicos dinâmicos.

5.8. Integração com Bases de Conhecimento Clínico

A qualidade de um agente de IA na área da saúde depende muito das informações que ele consegue acessar. Quanto melhores e mais completas forem essas fontes, mais útil ele será na prática. Essas informações ficam organizadas em bases de conhecimento clínico. Elas funcionam como grandes repositórios que guardam o saber médico acumulado ao longo do tempo. Ali estão dados sobre diagnósticos, tratamentos, medicamentos e protocolos de atendimento. Tudo isso é estruturado de forma que os sistemas consigam entender e usar facilmente. Ou seja, não é só informação armazenada é informação pronta para ser consultada e aplicada. Quando essas bases são integradas ao funcionamento do agente, ele deixa de ser um sistema genérico e passa a atuar de forma mais especializada. Na prática, isso significa que o agente consegue interpretar melhor cada situação, tomar decisões mais informadas e oferecer respostas mais úteis no contexto clínico.

5.8.1. Ontologias e terminologias médicas padronizadas

A interoperabilidade em sistemas de saúde depende fortemente do uso de terminologias e ontologias médicas padronizadas, que funcionam como uma linguagem comum entre diferentes plataformas clínicas e administrativas. Entre os principais padrões utilizados destaca-se o SNOMED CT[17], uma ontologia clínica abrangente que organiza conceitos médicos e suas relações semânticas. O CID-11[18] é amplamente utilizado para padronização de diagnósticos, enquanto o LOINC[19] fornece nomenclaturas e códigos para exames laboratoriais, medições clínicas e observações médicas.

Quando agentes de IA incorporam essas terminologias em seus mecanismos de processamento, passam a compreender de maneira mais robusta a linguagem médica utilizada em prontuários, laudos e registros clínicos. Isso inclui o reconhecimento de sinônimos, abreviações e diferentes formas de representação de uma mesma condição clínica. Na prática, o sistema consegue identificar que expressões como “insuficiência cardíaca congestiva”, “ICC” ou outras variações semânticas correspondem ao mesmo conceito médico subjacente.

Essa capacidade de normalização semântica torna os agentes mais robustos, interoperáveis e confiáveis, especialmente em ambientes clínicos reais, nos quais a documentação médica frequentemente apresenta elevada variabilidade terminológica e diferenças na forma de registro das informações. Além disso, o uso dessas ontologias facilita integração entre sistemas heterogêneos, melhora a recuperação de informações clínicas e contribui para maior consistência em processos de apoio à decisão médica baseados em inteligência artificial.

[17] <https://www.snomed.org/snomed-ct>
[18] <https://icd.who.int/en>
[19] <https://loinc.org>

5.8.2. Bases de literatura científica e diretrizes clínicas

Além do uso de ontologias médicas, agentes de IA aplicados à saúde tornam-se significativamente mais úteis quando conseguem acessar informações atualizadas provenientes da literatura científica biomédica. Um dos principais exemplos é o PubMed[20], uma das maiores bases de dados de literatura biomédica do mundo, mantida pela *National Library of Medicine*. Também existem plataformas especializadas em suporte à decisão clínica baseada em evidências, como o UpToDate[21] e o DynaMed[22], que organizam informações científicas na forma de recomendações clínicas estruturadas.

Quando agentes inteligentes se conectam a essas fontes utilizando técnicas como *Retrieval-Augmented Generation* (RAG), deixam de atuar apenas como sistemas geradores de respostas genéricas e passam a recuperar informações científicas confiáveis para fundamentar suas recomendações. Dessa forma, o agente pode não apenas responder uma pergunta clínica, mas também indicar artigos científicos, diretrizes médicas e níveis de evidência que sustentam determinada conduta terapêutica ou diagnóstica.

Um exemplo prático dessa abordagem é o Consensus[23], um mecanismo de busca baseado em IA voltado para literatura científica revisada por pares. O sistema permite que usuários realizem perguntas em linguagem natural e recebam respostas fundamentadas em artigos científicos, acompanhadas de resumos automáticos sobre o consenso existente na literatura acadêmica.

Na prática, esse tipo de integração auxilia médicos, pesquisadores e profissionais de saúde a economizar tempo durante processos de revisão bibliográfica e tomada de decisão clínica. Além disso, agentes capazes de recuperar evidências científicas atualizadas tendem a produzir recomendações mais transparentes, auditáveis e alinhadas às práticas da medicina baseada em evidências. No futuro, espera-se que sistemas desse tipo desempenhem papel cada vez mais relevante no suporte à decisão clínica, oferecendo assistência contextualizada e continuamente atualizada conforme a evolução da literatura biomédica.

5.8.3. Prontuários Eletrônicos e interoperabilidade via HL7 FHIR

A integração com Sistemas de Registro Eletrônico de Saúde (*Electronic Health Records* – EHR) ou Prontuários Eletrônicos do Paciente (PEP) representa um dos elementos centrais para viabilizar o uso prático de agentes de IA no ambiente clínico. Esses sistemas concentram informações essenciais sobre os pacientes, incluindo histórico médico, exames laboratoriais, diagnósticos, prescrições e evoluções clínicas. Para que diferentes plataformas consigam interoperar de maneira padronizada, utiliza-se amplamente o padrão HL7 FHIR (*Fast Healthcare Interoperability Resources*)[24], desenvolvido pela *Health Level Seven International*.

O padrão FHIR define uma estrutura organizada para representação e troca de dados clínicos, permitindo modelar recursos relacionados a pacientes, exames, diagnósticos,

[20] <https://pubmed.ncbi.nlm.nih.gov>
[21] <https://www.uptodate.com>
[22] <https://www.dynamed.com>
[23] <https://consensus.app>
[24] <https://www.hl7.org/fhir/>

prescrições, procedimentos e observações médicas. Na prática, isso possibilita que agentes inteligentes acessem informações clínicas por meio de APIs padronizadas, funcionando como uma interface segura e interoperável entre sistemas hospitalares e aplicações baseadas em inteligência artificial. Além da leitura de dados, esses agentes também podem atualizar registros clínicos, registrar evoluções médicas e auxiliar fluxos operacionais de forma integrada ao ecossistema hospitalar.

Diversas soluções comerciais já utilizam esse modelo de integração. Ferramentas como o Nabla Copilot[25] e o Abridge[26] conectam-se diretamente a sistemas hospitalares para auxiliar processos de documentação clínica. O Abridge, por exemplo, já é utilizado em instituições como o UPMC (*University of Pittsburgh Medical Center*)[27], nos Estados Unidos. Esses agentes conseguem recuperar automaticamente informações do paciente, processar transcrições de consultas e gerar anotações clínicas estruturadas em tempo real, salvando os dados diretamente no prontuário eletrônico.

Esse tipo de automação reduz significativamente o volume de tarefas administrativas executadas manualmente pelos profissionais de saúde, diminuindo erros de documentação e aumentando eficiência operacional. Como consequência, médicos e equipes clínicas podem dedicar mais tempo ao cuidado direto do paciente, reduzindo sobrecarga burocrática e melhorando a qualidade da assistência prestada.

5.9. Exemplos Reais: Aplicações de Referência Mundial

Esta seção apresenta os principais exemplos reais de agentes e sistemas de inteligência artificial aplicados à saúde em escala global. Cada caso demonstra como os fundamentos teóricos discutidos anteriormente se materializam em produtos e plataformas que já impactam pacientes, profissionais de saúde e instituições hospitalares em diferentes partes do mundo.

5.9.1. Ada Health: Triagem Clínica Baseada em IA

A Ada Health[28] foi fundada em 2011 em Berlim, Alemanha, e tornou-se uma das plataformas de avaliação de sintomas por inteligência artificial mais reconhecidas mundialmente. Seu principal produto consiste em um aplicativo de saúde disponível em mais de 130 países, capaz de conduzir entrevistas clínicas adaptativas com usuários e gerar listas priorizadas de possíveis condições médicas com base nos sintomas relatados.

O mecanismo central da plataforma combina raciocínio probabilístico bayesiano com modelos de processamento de linguagem natural. Diferentemente de árvores de decisão estáticas, o sistema adapta dinamicamente as próximas perguntas conforme as respostas fornecidas pelo usuário, aproximando-se do raciocínio clínico utilizado por médicos durante consultas reais. A base de conhecimento da plataforma cobre mais de 3.000 condições médicas, incluindo doenças raras, sendo continuamente revisada e atualizada por equipes médicas especializadas.

[25] <https://www.nabla.com/copilot>
[26] <https://www.abridge.com>
[27] <https://www.upmc.com>
[28] <https://ada.com>

Estudos publicados pela empresa e por pesquisadores independentes indicam que a Ada inclui o diagnóstico correto entre as três primeiras sugestões em mais de 90% dos casos avaliados, desempenho considerado comparável ao de médicos generalistas em cenários de triagem clínica. A plataforma também foi integrada a iniciativas de saúde pública em países africanos, como Ruanda e Tanzânia, com o objetivo de ampliar acesso à avaliação clínica em regiões com escassez de profissionais de saúde.

No Brasil, a empresa possui parcerias com operadoras de saúde e trabalha na adaptação cultural e linguística da plataforma ao contexto epidemiológico nacional, incluindo condições endêmicas como dengue, leishmaniose e doença de Chagas. Seu modelo de negócios baseado em integração B2B permite que hospitais e planos de saúde incorporem a tecnologia em fluxos digitais de triagem, reduzindo atendimentos desnecessários em pronto-socorros e direcionando pacientes para níveis adequados de cuidado.

5.9.2. Google DeepMind: AlphaFold e Gemini na Medicina

O Google DeepMind[29] produziu alguns dos avanços mais relevantes da inteligência artificial aplicada à saúde na última década, destacando-se especialmente o AlphaFold e os modelos multimodais Gemini voltados para aplicações biomédicas.

5.9.2.1. AlphaFold: Revolução na Biologia Estrutural

O AlphaFold[30] foi apresentado originalmente em 2020 e solucionou um dos maiores desafios da biologia molecular: a predição da estrutura tridimensional de proteínas a partir de sequências de aminoácidos. Antes dessa tecnologia, determinar experimentalmente a estrutura de proteínas podia demandar anos de trabalho e custos extremamente elevados.

O modelo utiliza arquiteturas derivadas de Transformers aplicadas ao processamento de sequências proteicas, permitindo prever estruturas com precisão próxima à experimental em poucos minutos. As implicações médicas são profundas, especialmente no contexto de doenças relacionadas a proteínas mal dobradas, como Alzheimer, Parkinson e diversos tipos de câncer.

Em parceria com o EMBL-EBI[31], o DeepMind disponibilizou publicamente um banco contendo mais de 200 milhões de estruturas proteicas previstas por IA. Empresas farmacêuticas como AstraZeneca, Bayer e Pfizer passaram a utilizar a tecnologia em processos de descoberta de medicamentos, enquanto a Isomorphic Labs[32] utiliza variantes do modelo para desenvolvimento de terapias baseadas em IA.

Em 2024, foi lançado o AlphaFold 3, expandindo as capacidades do sistema para modelagem de interações entre proteínas, DNA, RNA e pequenas moléculas terapêuticas.

[29] <https://deepmind.google>
[30] <https://alphafold.ebi.ac.uk>
[31] <https://www.ebi.ac.uk>
[32] <https://www.isomorphiclabs.com>

5.9.2.2. Med-PaLM 2 e Gemini: LLMs para Aplicações Médicas

O Med-PaLM 2[33] foi desenvolvido pela Google Research especificamente para aplicações médicas. Em avaliações padronizadas como o USMLE (*United States Medical Licensing Examination*), o modelo alcançou desempenho comparável ao de médicos especialistas, representando um marco importante na aplicação de LLMs em domínios clínicos altamente especializados.

O modelo demonstrou capacidade de responder perguntas abertas sobre casos clínicos, interpretar exames laboratoriais e gerar explicações avaliadas por médicos como úteis ou muito úteis. Posteriormente, o Gemini[34] expandiu essas capacidades ao introduzir processamento multimodal integrado de texto, imagens radiológicas, lâminas histopatológicas e imagens dermatológicas.

Estudos preliminares demonstraram desempenho competitivo na interpretação de radiografias de tórax e classificação de lesões dermatológicas. Além disso, a integração com mecanismos de busca e recuperação de informações científicas via RAG permite acessar literatura médica atualizada em tempo real, reduzindo limitações relacionadas ao corte temporal de treinamento dos modelos.

5.9.3. Infermedica: Plataforma de Triagem Probabilística

A Infermedica[35], fundada na Polônia em 2012, desenvolveu uma das plataformas de triagem clínica baseada em IA mais amplamente validadas internacionalmente. Diferentemente de sistemas baseados em árvores de decisão fixas, a plataforma utiliza modelos probabilísticos bayesianos capazes de avaliar centenas de condições clínicas, milhares de sintomas e múltiplos fatores de risco em dezenas de idiomas.

O paciente descreve seus sintomas iniciais e o sistema conduz entrevistas dinâmicas adaptativas em tempo real. Ao final, os casos são classificados em diferentes níveis de urgência clínica, incluindo emergência, urgência, semiurgência, não urgência e autocuidado. Estudos reportados pela empresa indicam elevada segurança clínica nas recomendações emitidas e redução significativa no tempo médio de triagem.

5.9.4. DAX Copilot (Nuance/Microsoft): Documentação Clínica Automatizada

O DAX Copilot[36] representa uma das aplicações mais avançadas de IA para documentação clínica automatizada. Desenvolvido originalmente pela Nuance Communications, posteriormente adquirida pela Microsoft em 2022, o sistema utiliza tecnologia de *ambient listening* para capturar automaticamente conversas entre médicos e pacientes durante consultas clínicas.

Após o término da consulta, o sistema gera automaticamente notas estruturadas no formato SOAP (*Subjective, Objective, Assessment, Plan*), integrando-as diretamente ao prontuário eletrônico. O sistema possui integração com plataformas como Epic EHR[37],

[33] <https://sites.research.google/med-palm/>
[34] <https://deepmind.google/technologies/gemini/>
[35] <https://infermedica.com>
[36] <https://www.microsoft.com/en-us/health-solutions/dragon-copilot>
[37] <https://www.epic.com>

amplamente utilizadas em hospitais norte-americanos.

Resultados publicados indicam redução significativa no tempo gasto com documentação médica, economia de vários minutos por consulta e diminuição importante nos índices de burnout associados à sobrecarga administrativa. Em versões mais recentes, o sistema passou a incluir funcionalidades adicionais, como sugestão automática de códigos de faturamento e identificação de lacunas em cuidados preventivos.

5.9.5. Tempus AI: Medicina de Precisão em Oncologia

A Tempus AI[38], fundada em Chicago em 2015, opera uma das maiores plataformas de medicina de precisão voltadas à oncologia. A empresa integra milhões de registros clínicos com dados de sequenciamento genômico avançado, incluindo exoma completo e RNA-seq, criando bases de dados utilizadas para desenvolvimento de modelos preditivos e suporte terapêutico personalizado.

A plataforma identifica mutações tumorais relevantes, sugere terapias-alvo aprovadas pela FDA associadas às alterações genéticas encontradas e identifica ensaios clínicos compatíveis com o perfil molecular do paciente. Esse modelo auxilia oncologistas na identificação de opções terapêuticas especialmente em cânceres raros ou casos complexos.

Além disso, a empresa desenvolve modelos capazes de prever resposta terapêutica, risco de progressão tumoral e probabilidade de recorrência da doença, permitindo estratégias mais individualizadas de acompanhamento clínico.

5.9.6. Abridge: Transcrição e Sumarização de Consultas

A Abridge[39] é uma startup norte-americana fundada em 2018 voltada à transcrição e sumarização automática de consultas médicas utilizando inteligência artificial. O sistema captura conversas entre médicos e pacientes e gera automaticamente resumos estruturados contendo diagnósticos discutidos, medicamentos prescritos, exames solicitados e planos terapêuticos.

A plataforma é utilizada pelo UPMC (*University of Pittsburgh Medical Center*), um dos maiores sistemas hospitalares dos Estados Unidos. Estudos preliminares indicam redução significativa da carga de documentação pós-consulta e aumento da satisfação dos profissionais com a qualidade das notas clínicas produzidas automaticamente.

O sistema foi desenvolvido especificamente para lidar com a elevada variabilidade linguística observada em consultas reais, incluindo terminologia médica, abreviações, interrupções e mudanças frequentes de contexto.

5.9.7. Hospital Israelita Albert Einstein: HStory

No contexto brasileiro, o Hospital Israelita Albert Einstein[40] desenvolveu o HStory, uma plataforma baseada em IA generativa voltada à consolidação inteligente de dados clínicos distribuídos em múltiplos sistemas hospitalares.

[38]<https://www.tempus.com>
[39]<https://www.abridge.com>
[40]<https://www.einstein.br>

A solução foi criada para enfrentar um problema recorrente em hospitais de grande porte: a fragmentação das informações clínicas ao longo de diferentes atendimentos, especialidades e sistemas de informação. O HStory utiliza agentes inteligentes para reunir, estruturar e sintetizar automaticamente essas informações em um painel clínico centralizado acessível durante o atendimento médico.

Dessa forma, em vez de navegar manualmente por múltiplos registros fragmentados, o profissional recebe uma narrativa clínica consolidada contendo histórico de internações, diagnósticos prévios, alergias, medicamentos em uso, exames relevantes e procedimentos anteriores. A iniciativa é considerada uma das aplicações mais avançadas de IA generativa em ambiente hospitalar na América Latina.

5.10. Benefícios dos Agentes de IA na Saúde

A incorporação de agentes inteligentes em ambientes clínicos e hospitalares representa uma das transformações mais profundas vivenciadas pelo setor de saúde nas últimas décadas. Impulsionados pelos avanços em aprendizado de máquina, processamento de linguagem natural, processamento de imagens e sistemas especialistas, esses agentes têm demonstrado capacidade de atuar em múltiplas frentes operacionais, desde a burocracia administrativa até o suporte direto ao raciocínio diagnóstico. Esta seção apresenta os principais benefícios documentados na literatura científica, organizados em cinco eixos temáticos como automação de processos clínicos, apoio à decisão médica, redução de sobrecarga administrativa, acesso à literatura científica e personalização no atendimento ao paciente.

5.10.1. Automação de Processos Clínicos

A automação de processos clínicos constitui um dos eixos de maior impacto imediato da inteligência artificial na saúde. Processos que historicamente demandavam horas de trabalho humano como o agendamento de consultas, a triagem de pacientes, o monitoramento de sinais vitais e a reconciliação de medicamentos passam a ser executados por agentes automáticos com maior velocidade, menor taxa de erro e disponibilidade contínua ao longo de todas as horas do dia (TOPOL, 2019).

Em ambientes de pronto-socorro, por exemplo, agentes inteligentes são capazes de realizar a triagem inicial dos pacientes com base em sintomas relatados, histórico clínico e dados de sinais vitais coletados. Estudos demonstram que sistemas de triagem baseados em IA apresentam acurácia comparável ou superior à triagem humana convencional em categorias de risco intermediário, além de reduzir significativamente o tempo de espera (ALOMARI et al., 2025). Esta capacidade não substitui o julgamento médico, mas otimiza o fluxo de atendimento, redirecionando casos críticos mais rapidamente para o cuidado especializado.

Outra área de aplicação relevante é o monitoramento contínuo de pacientes internados. Sensores conectados a plataformas de IA permitem o acompanhamento em tempo real de parâmetros como frequência cardíaca, saturação de oxigênio e pressão arterial, com geração automática de alertas quando valores se afastam de limiares pré-estabelecidos. Sistemas como o de detecção de sepse baseado em IA têm demonstrado redução de mortalidade em ambientes de UTI, ao permitir intervenções antes que o quadro se torne crítico

(YUAN et al., 2020).

O processo de verificação e atualização da lista de medicamentos de um paciente ao longo das transições de cuidado é uma etapa essencial para a segurança clínica e tem se beneficiado significativamente da automação. Isso é especialmente importante quando se considera que erros de medicação estão entre as causas mais frequentes de eventos adversos em ambientes hospitalares. Nesse cenário, agentes de inteligência artificial integrados a prontuários eletrônicos têm se mostrado ferramentas estratégicas. Esses sistemas conseguem analisar prescrições em tempo real, identificando incompatibilidades, duplicidades e possíveis erros de dosagem antes mesmo que o medicamento seja administrado ao paciente.

Estudos sobre sistemas de suporte à decisão clínica demonstram que essas tecnologias contribuem de forma significativa para a redução de erros de medicação e para o aumento da segurança do paciente (SUTTON et al., 2020).

5.10.2. Apoio à Decisão Médica

O apoio à decisão médica representa talvez o benefício mais complexo e transformador da IA clínica. Diferentemente da automação de tarefas rotineiras, os sistemas de suporte à decisão (Clinical Decision Support Systems — CDSS) atuam diretamente na esfera do raciocínio diagnóstico e terapêutico, análise de exames e formulação de hipóteses clínicas. Seu objetivo central não é substituir o médico, mas ampliar a capacidade cognitiva do profissional diante de cenários de alta complexidade e volume massivo de dados (OBERMEYER; EMANUEL, 2016). Modelos baseados em redes neurais profundas (deep learning) vêm apresentando resultados expressivos na análise de imagens médicas. Na dermatologia, por exemplo, sistemas de inteligência artificial já conseguem identificar melanomas malignos com sensibilidade comparável à de dermatologistas experientes, a partir de imagens dermatoscópicas (ESTEVA et al., 2017).

Na radiologia, modelos voltados à detecção de nódulos pulmonares em tomografias computadorizadas alcançam altas taxas de verdadeiro-positivo, frequentemente superiores a 94%, mantendo baixos índices de falsos positivos e contribuindo para o diagnóstico precoce do câncer de pulmão. De forma semelhante, na oftalmologia, algoritmos aplicados à análise de imagens de fundo de olho têm demonstrado elevada precisão na classificação de doenças como retinopatia diabética e degeneração macular, em alguns casos superando o desempenho de especialistas (GULSHAN et al., 2016).

Para além da análise de imagens, os sistemas de IA também se destacam pela capacidade de integrar diferentes tipos de dados clínicos, como informações laboratoriais, genômicas, farmacológicas e históricas do paciente, a fim de apoiar decisões terapêuticas mais personalizadas. No contexto da oncologia de precisão, plataformas como o *Watson for Oncology*[41], desenvolvido pela IBM, assim como sistemas acadêmicos criados em centros de pesquisa, têm demonstrado alta concordância com diretrizes clínicas consolidadas, chegando a recomendar tratamentos consistentes em até 96% dos casos analisados (SOMASHEKHAR et al., 2018).

Esse desempenho está diretamente relacionado à capacidade desses sistemas de

[41] https://www.ibm.com/mysupport/s/topic/0TO500000002PWlGAM/watson-for-oncology?language=en_US

analisar simultaneamente um grande volume de variáveis clínicas algo que, na prática, ultrapassa os limites do raciocínio humano sem suporte computacional. Outro ponto relevante diz respeito à redução de vieses cognitivos presentes no processo de decisão clínica. Tendências como o viés de disponibilidade em que diagnósticos recentes influenciam desproporcionalmente novas avaliações ou o fechamento prematuro do diagnóstico podem comprometer a qualidade da análise médica. Nesse sentido, sistemas de apoio à decisão clínica contribuem ao sugerir hipóteses alternativas baseadas em padrões estatísticos e dados populacionais, incentivando uma abordagem mais abrangente e menos suscetível a erros sistemáticos (GRABER, 2013).

5.10.3. Redução de Sobrecarga Administrativa

A sobrecarga administrativa representa um dos principais fatores de esgotamento profissional entre médicos e enfermeiros em todo o mundo. Estima-se que profissionais de saúde nos Estados Unidos dedicam entre 35% e 50% de sua jornada de trabalho a tarefas burocráticas, como documentação em prontuários, preenchimento de formulários de autorização, codificação de procedimentos e correspondência administrativa tempo que poderia ser direcionado ao cuidado direto ao paciente (SINSKY et al., 2016).

Agentes de processamento de linguagem natural (PLN) oferecem uma solução direta para esse problema, sendo capazes de transcrever consultas médicas em tempo real, estruturar automaticamente o prontuário eletrônico e codificar procedimentos nos padrões internacionais como CID-10 e SNOMED-CT. Ferramentas de documentação ambiental, como Nuance DAX e Suki AI, têm sido adotadas em sistemas de saúde norte-americanos e europeus, com evidências de redução significativa no tempo de documentação clínica e na carga de trabalho dos profissionais. Estudos apontam diminuições na ordem de 20% a 30% no tempo gasto com registros, além da redução de horas extras dedicadas à finalização de prontuários (HABERLE et al., 2024; RAZAGHI et al., 2026).

No contexto brasileiro, o Sistema Único de Saúde (SUS) enfrenta desafios adicionais relacionados ao volume de processos administrativos em instituições com recursos limitados. A implementação de agentes automáticos para triagem de solicitações de exames, autorização de internações e emissão de laudos tem potencial de agilizar significativamente o fluxo assistêencial, reduzindo filas e o tempo de espera para procedimentos eletivos (Brasil, Ministério da Saúde, 2022). As principais áreas de redução de carga administrativa por meio de agentes de IA incluem:

- Transcrição automática de consultas e geração de prontuários estruturados;
- Codificação automática de diagnósticos e procedimentos (CID-10, TUSS, SNOMED-CT);
- Processamento e roteamento inteligente de solicitações de autorização de planos de saúde;
- Geração automática de relatórios e sumários de alta hospitalar;
- Agendamento e reescalonamento inteligente de consultas e exames.

O impacto da redução administrativa vai além da eficiência operacional estudos apontam correção direta entre menor carga burocrática e menor taxa de burnout entre profissionais de saúde, com reflexo na qualidade do cuidado prestado e na satisfação dos pacientes (LINZER et al., 2015).

5.10.4. Acesso Rápido à Literatura Científica

O volume de publicações científicas na área médica cresce de forma exponencial. Em 2023, a base de dados PubMed indexou mais de 36 milhões de artigos, com cerca de 1,3 milhão de novos trabalhos adicionados anualmente. Diante desse cenário, torna-se humanamente impossível para qualquer profissional da saúde manter-se atualizado em todas as subáreas relevantes para sua prática, criando uma lacuna entre o conhecimento científico disponível e a prática clínica cotidiana (BASTIAN; GLASZIOU; CHALMERS, 2010).

Agentes de IA baseados em modelos de linguagem de grande escala (LLMs) oferecem uma resposta concreta a esse desafio. Sistemas como o PubMedBERT[42], BioGPT[43] e ferramentas integradas a plataformas clínicas são capazes de realizar buscas na literatura, identificar evidências relevantes para uma pergunta clínica específica e sintetizar as principais conclusões de múltiplos estudos em linguagem natural, respeitando os princípios da medicina baseada em evidências (SINGHAL et al., 2023).

Uma das aplicações mais promissoras é a geração automática de revisões sistemáticas assistidas, nas quais o agente realiza a triagem de artigos, extrai dados padronizados de cada estudo e organiza os resultados em formatos compatíveis com as diretrizes PRISMA. Embora a validação humana continue indispensável, o auxílio automatizado reduz em semanas o tempo necessário para a conclusão de uma revisão abrangente (MARSHALL; WALLACE, 2019). No ambiente clínico, a integração de agentes de busca literária ao prontuário eletrônico permite que, no momento do atendimento, o profissional receba sugestões contextualizadas de diretrizes clínicas, meta-análises e protocolos baseados em evidências relacionados ao caso em tela.

5.10.5. Personalização no Atendimento ao Paciente

A medicina personalizada também denominada medicina de precisão parte do princípio de que cada paciente é único em sua constituição genômica, história clínica, estilo de vida e contexto social. Agentes de IA viabilizam, em escala, a aplicação desse paradigma, integrando dados heterogêneos de múltiplas fontes para construir perfis individuais e gerar recomendações altamente adaptadas a cada cenário particular (COLLINS; VARMUS, 2015). No âmbito da farmacogenômica, agentes de IA analisam o perfil genético do paciente para prever respostas a medicamentos específicos, identificar riscos de reações adversas e recomendar ajustes de dosagem individualizados. Estudos em oncologia demonstram que o tratamento guiado por análise genômica assistida por IA resulta em taxas de resposta superior à quimioterapia convencional não-dirigida, com menor toxicidade e melhor qualidade de vida (KRZYSZCZYK et al., 2018).

A personalização também envolve a forma de comunicação e o acompanhamento ao longo do tempo. Assistentes virtuais clínicos têm sido usados para monitorar pacientes

[42]

[43]https://the-decoder.com/biogpt-is-a-microsoft-language-model-trained-for-biomedical-tasks/

com doenças crônicas fora do hospital, realizando contatos periódicos, identificando sinais iniciais de piora e incentivando o uso correto dos medicamentos por meio de lembretes e orientações adaptadas a cada paciente. Estudos com pessoas com diabetes e hipertensão mostram que o uso de chatbots clínicos pode melhorar a adesão ao tratamento e os resultados de saúde após seis meses (SHAN; SARKAR; MARTIN, 2019). Essa abordagem também aparece na adaptação da linguagem e do canal de comunicação às preferências do paciente.

Sistemas de IA conseguem identificar o nível de letramento em saúde do usuário e ajustar automaticamente a complexidade das informações, o que facilita a compreensão e aumenta o engajamento. Em populações com baixo letramento funcional comum em contextos de baixa renda essa capacidade pode contribuir diretamente para reduzir desigualdades no acesso à informação em saúde (MK, 2007). Por fim, essa abordagem orientada por agentes também se estende ao campo da prevenção. Com base em padrões de comportamento, dados de dispositivos vestíveis e histórico clínico, sistemas de IA podem elaborar planos preventivos mais ajustados a cada indivíduo, identificando fatores de risco e sugerindo mudanças de hábitos antes do surgimento de doenças crônicas. Essa mudança de foco do tratamento para a prevenção tem potencial para reduzir significativamente os custos em saúde no longo prazo, além de favorecer uma vida mais longa e com melhor qualidade (BHAVNANI; NARULA; SENGUPTA, 2016).

5.10.6. Bases de dados farmacológicos e de interações medicamentosas

Outro ponto muito importante nessa integração com bases de conhecimento é o acesso a informações sobre medicamentos. Existem bases como DrugBank, RxNorm e Micromedex que reúnem dados detalhados sobre milhares de remédios, incluindo para que servem, quando não devem ser usados e como interagem entre si.

Quando agentes de IA têm acesso a essas bases, eles conseguem atuar diretamente no momento da prescrição. Por exemplo, ao analisar os medicamentos que um paciente já utiliza, o agente pode identificar automaticamente possíveis interações perigosas e alertar o profissional de saúde antes que o problema aconteça. Isso complementa o que vimos antes: o agente não só utiliza evidências científicas (via RAG) e dados do paciente (via EHR), mas também aplica conhecimento farmacológico específico para aumentar a segurança do tratamento. Na prática, esse tipo de solução já está sendo utilizado.

A Santa Casa de Belo Horizonte implementou um sistema em que um agente de IA acompanha continuamente as prescrições médicas. Ele cruza essas informações com bases farmacológicas e gera alertas organizados por nível de risco. O resultado é um cuidado mais proativo: em vez de identificar problemas depois que eles acontecem, o sistema ajuda a prevenir eventos adversos, destacando casos críticos com antecedência — algo que, no modelo manual, poderia passar despercebido ou ser detectado tarde demais.

5.11. Desafios e Limitações

Apesar do potencial transformador demonstrado nas seções anteriores, a adoção de agentes de IA na saúde enfrenta desafios técnicos, éticos, regulatórios e operacionais significativos. A compreensão dessas limitações é essencial para garantir desenvolvimento responsável, segurança clínica e implementação sustentável dessas tecnologias em ambientes hospitalares

e assistenciais.

5.11.1. Riscos de Alucinação em Modelos Generativos

O fenômeno das chamadas "alucinações" situações nas quais modelos generativos produzem informações factualmente incorretas apresentadas com aparente confiança representa um dos riscos mais críticos da aplicação de LLMs em saúde. Um modelo pode sugerir dosagens incorretas de medicamentos, inventar referências bibliográficas inexistentes ou recomendar condutas contraindicadas utilizando linguagem altamente convincente. Em ambientes clínicos, nos quais decisões equivocadas podem gerar danos diretos aos pacientes, essa limitação exige mecanismos robustos de mitigação. Entre as estratégias mais utilizadas destacam-se arquiteturas baseadas em *Retrieval-Augmented Generation* (RAG) com fontes controladas, camadas adicionais de validação, implementação de *guardrails* automatizados e supervisão humana obrigatória em decisões clínicas de maior impacto.

Estudos recentes demonstram que mesmo modelos considerados estado da arte, como GPT-4 e Med-PaLM 2, apresentam taxas mensuráveis de alucinação em tarefas médicas específicas. Nesse contexto, abordagens híbridas de mitigação têm se mostrado mais eficazes, incluindo recuperação de informações em bases verificadas, utilização de parâmetros de geração mais conservadores, instruções explícitas para indicação de incerteza e avaliação sistemática utilizando frameworks especializados como RAGAS[44] e TruLens[45].

5.11.2. Segurança de Dados e Conformidade com a LGPD

Dados de saúde figuram entre as informações mais sensíveis existentes. A integração de agentes inteligentes com prontuários eletrônicos, sistemas laboratoriais e bases genômicas levanta questões críticas relacionadas ao armazenamento, transmissão e processamento seguro dessas informações.

No Brasil, a Lei Geral de Proteção de Dados (LGPD)[46] classifica dados de saúde como dados pessoais sensíveis, impondo requisitos rigorosos para tratamento, armazenamento e compartilhamento dessas informações. Entre esses requisitos destacam-se necessidade de base legal adequada, princípios de minimização de dados, transparência no tratamento e garantia de segurança da informação. Vazamentos de dados médicos possuem consequências particularmente graves, podendo expor condições clínicas sensíveis, comprometer privacidade de pacientes e gerar impactos sociais, financeiros e psicológicos relevantes.

Além disso, órgãos reguladores como a ANVISA[47] e o Conselho Federal de Medicina (CFM)[48] têm avançado na definição de diretrizes específicas para sistemas de IA com aplicação clínica, incluindo requisitos relacionados à validação clínica, rastreabilidade de decisões e responsabilização institucional. Nesse cenário, conformidade regulatória deixa de ser apenas uma exigência jurídica e passa a representar elemento central para

[44] <https://github.com/explodinggradients/ragas>
[45] <https://www.trulens.org>
[46] <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>
[47] <https://www.gov.br/anvisa/pt-br>
[48] <https://portal.cfm.org.br>

confiança, segurança e adoção sustentável dessas tecnologias.

5.11.3. Viés Algorítmico e Equidade em Saúde

Modelos de inteligência artificial aprendem padrões a partir de dados históricos, os quais frequentemente refletem desigualdades estruturais presentes nos sistemas de saúde. Como consequência, algoritmos treinados predominantemente com determinadas populações podem apresentar desempenho inferior quando aplicados a grupos sub-representados no conjunto de treinamento.

Estudos demonstram que sistemas de predição de risco clínico podem apresentar diferenças significativas de desempenho relacionadas a raça, gênero, faixa etária e condição socioeconômica. Um caso amplamente discutido na literatura mostrou que um algoritmo utilizado nos Estados Unidos atribuía sistematicamente menor prioridade de cuidado a pacientes negros em comparação a pacientes brancos com gravidade clínica semelhante, perpetuando desigualdades já existentes no sistema de saúde.

No contexto brasileiro, essa questão torna-se particularmente relevante devido à elevada heterogeneidade populacional e às profundas desigualdades regionais e socioeconômicas. Dessa forma, o desenvolvimento de datasets representativos da população brasileira, aliado à validação específica em grupos vulneráveis, constitui requisito fundamental para adoção ética e segura de IA em saúde pública.

5.11.4. Limites da Automação e Responsabilidade Profissional

A crescente automação de processos clínicos levanta questões fundamentais relacionadas à responsabilidade profissional e ao papel do julgamento humano na tomada de decisão médica. Segundo diretrizes do Conselho Federal de Medicina, a responsabilidade final pela conduta clínica permanece sempre sob responsabilidade do profissional médico, independentemente do suporte tecnológico utilizado. Isso implica que agentes inteligentes devem ser concebidos como ferramentas de apoio à decisão clínica, e não como substitutos do raciocínio médico. Além disso, profissionais de saúde precisam ser adequadamente treinados para compreender tanto capacidades quanto limitações dos sistemas baseados em IA.

Outro risco relevante é o fenômeno conhecido como "automação excessiva", no qual profissionais passam a aceitar recomendações automatizadas sem avaliação crítica adequada. Estudos relacionados à chamada *alert fatigue* demonstram que excesso de notificações e alertas pode levar usuários a ignorarem sistematicamente mensagens do sistema inclusive aquelas realmente importantes. Para mitigar esses problemas, tornam-se essenciais estratégias como design cuidadoso de interfaces, priorização inteligente de alertas, mecanismos de explicabilidade e construção de cultura institucional voltada ao uso crítico e supervisionado das ferramentas de IA.

5.11.5. Barreiras de Adoção e Infraestrutura

A adoção prática de agentes inteligentes na saúde também enfrenta barreiras estruturais significativas, especialmente em ambientes com recursos limitados. Sistemas baseados em IA dependem fortemente da qualidade, completude e padronização dos dados clínicos disponíveis. Entretanto, muitos hospitais ainda operam com prontuários parcialmente

digitalizados ou sistemas fragmentados sem interoperabilidade adequada.

Além disso, a infraestrutura computacional necessária para treinamento e execução de modelos de linguagem de larga escala pode representar custo elevado para instituições de saúde. A utilização de GPUs especializadas, armazenamento seguro de grandes volumes de dados e integração contínua com sistemas hospitalares demandam investimentos financeiros e técnicos substanciais.

Nesse contexto, arquiteturas híbridas têm emergido como alternativa pragmática. Essas abordagens combinam processamento local (*on-premises*) para dados mais sensíveis com utilização de serviços em nuvem para tarefas de maior demanda computacional, buscando equilibrar desempenho, custo operacional e privacidade das informações clínicas.

5.12. Implementação Estudo de Caso

Esta seção apresenta um estudo de caso clínico hipotético com o objetivo de ilustrar, de forma concreta e detalhada, como um sistema multi-agente de IA operaria na prática em um contexto de urgência e emergência. O cenário é fictício, mas clinicamente plausível, construído com base em protocolos reais de atendimento de urgência.

5.12.1. Visão Geral da Arquitetura do Sistema

O sistema proposto é composto por três agentes especializados que operam de forma coordenada, cada um com escopo e responsabilidades bem definidos. O Agente de Triagem é responsável pela recepção da queixa inicial, classificação de gravidade e encaminhamento interno. O Agente Especialista realiza o raciocínio diagnóstico aprofundado, recupera diretrizes clínicas relevantes via RAG e formula hipóteses e condutas fundamentadas em evidências. O Agente de Documentação captura o raciocínio clínico produzido ao longo da interação e gera automaticamente o registro estruturado no prontuário eletrônico.

A orquestração entre os agentes segue um grafo de decisão implementado em LangGraph. O Agente de Triagem atua como ponto de entrada, avalia o nível de urgência e determina se o caso deve ser resolvido localmente, encaminhado ao Agente Especialista ou escalado imediatamente para atendimento presencial de emergência. O Agente de Documentação opera em paralelo, monitorando toda a troca de informações e consolidando o registro ao final do atendimento. O médico mantém controle total em cada etapa, e o sistema nunca executa ações clínicas sem que haja confirmação ou ao menos ciência do profissional responsável.

5.12.2. Contexto do Cenário

O cenário envolve um clínico geral que atende em consultório particular em uma cidade de médio porte no interior do Piauí. O consultório dispõe de recursos diagnósticos básicos, incluindo oxímetro, aparelho de pressão, glicosímetro e eletrocardiógrafo portátil, além de acesso limitado a especialistas presenciais. O médico utiliza um sistema de prontuário eletrônico integrado à plataforma multiagente, acessível por interface web durante a consulta.

O paciente é um homem de 61 anos, hipertenso em uso de losartana 50 mg por dia e metformina 850 mg duas vezes ao dia para diabetes tipo 2, tabagista de 30 maços ano,

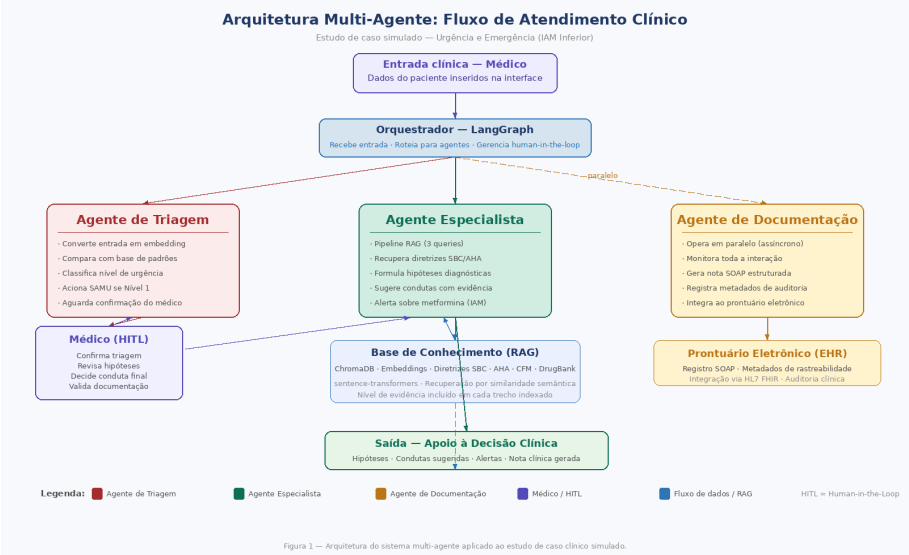

Figura 5.6. Diagrama Multi-agente

que chega ao consultório acompanhado da esposa às 10h47 com queixa de dor no peito de início súbito há aproximadamente 40 minutos, irradiando para o braço esquerdo, associada a sudorese fria e discreta falta de ar. O médico insere essas informações na interface do sistema e aciona o pipeline multiagente.

5.12.3. Etapa 1 Agente de Triagem

Ao receber a descrição clínica, o Agente de Triagem converte o texto em embedding e o compara com sua base de padrões de triagem indexados. Em menos de dois segundos, o agente identifica a combinação de dor torácica irradiada, sudorese e dispneia em paciente com perfil de risco cardiovascular elevado como um padrão de alta probabilidade para síndrome coronariana aguda (SCA). Com base nos protocolos do Manchester Triage System e nas diretrizes da Sociedade Brasileira de Cardiologia, o agente classifica o caso como Nível 1 Emergência Cardiovascular Provável, recomenda o acionamento imediato do SAMU (192) e solicita confirmação do médico antes de acionar o Agente Especialista. Simultaneamente, o sistema exibe um lembrete automatizado para checagem imediata de saturação de oxigênio, pressão arterial e realização do ECG portátil procedimentos iniciais recomendados enquanto o raciocínio diagnóstico aprofundado é processado em paralelo.

Além da classificação inicial, o agente também realiza uma análise contextual do histórico clínico previamente registrado no prontuário eletrônico. Informações como hipertensão arterial, diabetes mellitus, tabagismo, histórico familiar de infarto agudo do miocárdio e uso contínuo de medicamentos cardiovasculares são automaticamente incorporadas ao processo de priorização clínica. Essa etapa permite aumentar a sensibilidade da triagem e reduzir o risco de subestimação de casos críticos.

5.12.4. Etapa 2 Agente Especialista

Com a confirmação do médico, o Agente Especialista recebe o caso completo incluindo a classificação do Agente de Triagem, o histórico do paciente e os dados inseridos e

Figura 5.7. Agente de Triagem

inicia o pipeline RAG. O agente formula três queries semânticas para recuperação de documentos a primeira buscando diretrizes de manejo de SCA em atenção primária e consultório, a segunda focando em contraindicações e cuidados com antiagregação em diabéticos hipertensos e a terceira recuperando protocolos de estabilização pré-hospitalar para encaminhamento de emergência. O sistema recupera trechos relevantes da Diretriz Brasileira de Síndromes Coronarianas Agudas (SBC, 2021), do Protocolo de Suporte Avançado de Vida Cardiovascular da American Heart Association adaptado para cenários de baixa complexidade, e das orientações do CFM sobre responsabilidades do médico assistente em situações de emergência.

Com base nesses documentos todos com nível de evidência A, o Agente Especialista formula três hipóteses diagnósticas em ordem decrescente de probabilidade IAMCSST como hipótese principal, Angina Instável / IAMSST como segunda hipótese, e Dissecção Aórtica como diagnóstico diferencial a ser excluído. O agente sugere condutas de suporte enquanto o SAMU é acionado: AAS 200–300 mg VO mastigado (na ausência de contraindicação), nitroglicerina sublingual 0,5 mg se pressão arterial sistólica acima de 90 mmHg, oxigênio suplementar se SpO abaixo de 94%, acesso venoso periférico,

posicionamento semissentado e monitorização contínua. O agente emite ainda um alerta crítico a metformina em uso pelo paciente deve ser suspensa imediatamente diante da suspeita de IAM, dado o risco de acidose lática em contexto de hipoperfusão tecidual informação que poderia ser facilmente negligenciada na urgência do atendimento.

O médico revisa as sugestões, decide administrar AAS 300 mg mastigado e inicia o ECG portátil. O traçado evidencia supradesnivelamento de ST em DII, DIII e aVF, compatível com IAM inferior. O médico registra esse achado na interface e aciona o SAMU. O Agente Especialista, ao receber a atualização do ECG, complementa seu raciocínio identificando o padrão como sugestivo de oclusão da artéria coronária direita e recomenda antecipar o relato ao serviço de hemodinâmica referenciado o sistema recupera automaticamente o protocolo de ativação da sala de hemodinâmica do hospital de referência regional e o exibe ao médico.

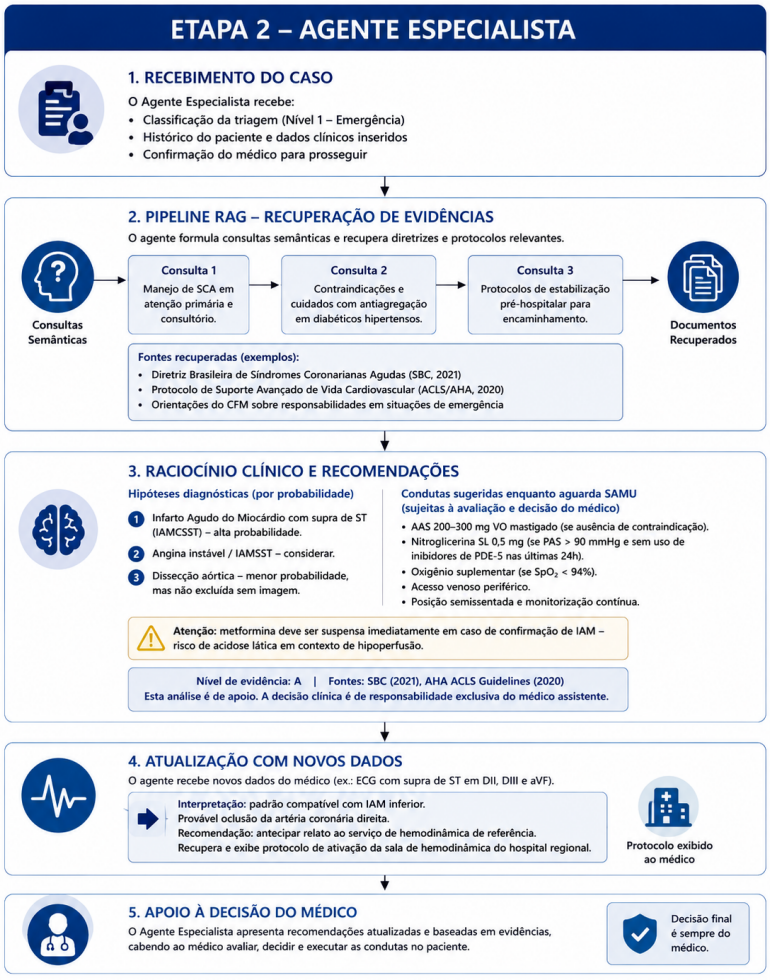

Figura 5.8. Agente Especialista

5.12.5. Etapa 3 Agente de Documentação

Enquanto o médico conduz o atendimento, o Agente de Documentação opera em segundo plano monitorando toda a interação. Ao término do atendimento, com a chegada do SAMU

e transferência do paciente, o agente consolida automaticamente um registro clínico estruturado no prontuário eletrônico, organizado no formato SOAP. O registro inclui informações subjetivas relacionadas à queixa e à história clínica, dados objetivos como sinais vitais e achados do ECG, avaliação contendo hipóteses diagnósticas e raciocínio clínico, além do plano terapêutico com medicamentos administrados, encaminhamentos realizados, serviço acionado e horários registrados.

O registro gerado também incorpora metadados de rastreabilidade, incluindo quais documentos foram recuperados pelo pipeline RAG, quais sugestões foram aceitas ou descartadas pelo médico e o horário correspondente a cada etapa do atendimento. Esse nível de documentação automatizada contribui para a continuidade do cuidado no serviço receptor, oferece suporte à auditoria clínica e produz evidências objetivas do raciocínio clínico empregado em eventuais análises jurídicas ou administrativas.

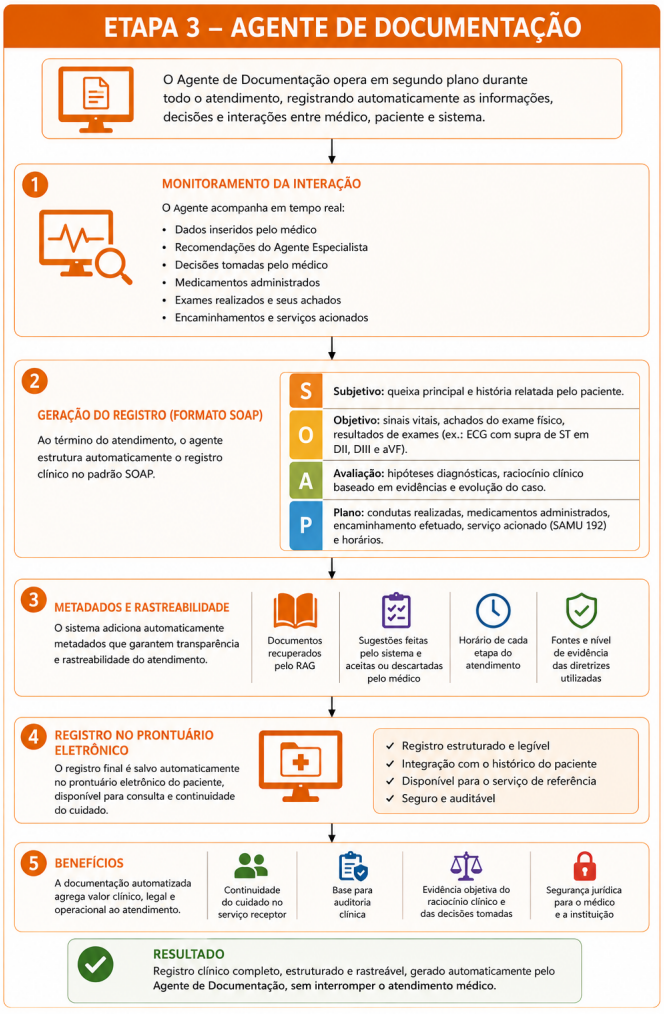

Figura 5.9. Agente de Documentação

5.12.6. Análise do Fluxo Multi-agente

O caso ilustra as vantagens da arquitetura multiagente em relação a soluções monolíticas. A especialização de responsabilidades entre triagem, raciocínio diagnóstico e documentação

permite que cada componente seja otimizado, atualizado e auditado de forma independente. Além disso, a paralelização entre o Agente Especialista e o Agente de Documentação reduz significativamente a latência operacional que existiria caso ambas as funções fossem executadas sequencialmente pelo mesmo sistema.

O princípio de *human in the loop* é preservado em todas as transições críticas do fluxo. O Agente de Triagem aguarda confirmação antes de acionar o Agente Especialista. As condutas sugeridas pelo Agente Especialista são apresentadas explicitamente como recomendações de apoio à decisão clínica. O Agente de Documentação registra não apenas as recomendações fornecidas pelo sistema, mas também as decisões efetivamente adotadas pelo profissional responsável. Essa distinção possui relevância ética, regulatória e jurídica.

Do ponto de vista do desempenho, o tempo total de processamento entre a inserção inicial dos dados e a apresentação do raciocínio clínico pelo Agente Especialista é estimado em menos de oito segundos em condições normais de conectividade. A documentação automatizada ocorre de forma assíncrona e não interfere diretamente na dinâmica do atendimento médico.

5.12.7. Limitações e Considerações do Cenário

É importante reconhecer as limitações inerentes ao estudo de caso apresentado. Por se tratar de um cenário simulado, os tempos de resposta, a qualidade dos documentos recuperados e o comportamento do sistema em situações limite, como perguntas fora do escopo da base de conhecimento ou dados clínicos contraditórios, ainda não foram empiricamente validados. A validação clínica formal do sistema, envolvendo especialistas e ambientes reais de atendimento, constitui etapa indispensável antes de qualquer implantação prática.

Além disso, o cenário pressupõe conectividade estável e integração com prontuário eletrônico dotado de API compatível, condições que nem sempre estão disponíveis em consultórios do interior do Brasil. Dessa forma, o desenvolvimento de modos de operação degradada, capazes de manter funcionalidades essenciais mesmo em ambientes de baixa conectividade, representa um requisito técnico importante para as próximas etapas do projeto.

5.13. Conclusão

Os agentes de Inteligência Artificial representam uma das fronteiras mais promissoras da computação aplicada à saúde. Ao combinar a capacidade generativa dos grandes modelos de linguagem com técnicas de recuperação de informação, integração padronizada com sistemas externos por meio de protocolos como o MCP, e arquiteturas multiagentes capazes de distribuir responsabilidades especializadas, esses sistemas abrem possibilidades inéditas para o suporte à decisão clínica, a automação de processos e a personalização do cuidado.

Ao longo deste minicurso, percorremos o caminho que vai dos fundamentos conceituais a definição de agente inteligente, a evolução histórica da IA na saúde, a distinção entre chatbots e agentes autônomos até os desafios técnicos, éticos e regulatórios que condicionam uma adoção responsável dessas tecnologias. Exploramos em profundidade exemplos reais de referência global, como a Ada Health com sua triagem probabilística adaptativa, o Google DeepMind com o AlphaFold e o Med-PaLM 2, a Infermedica com

sua plataforma multilíngue, o DAX Copilot com a automação de documentação clínica e a Tempus AI com a medicina de precisão em oncologia. No contexto brasileiro, o HStory do Hospital Israelita Albert Einstein demonstra que o país também produz inovação de alto nível nessa área.

A mensagem central que este material busca transmitir é de equilíbrio: entre o entusiasmo com o potencial transformador dessas tecnologias e a responsabilidade que seu uso em contextos de saúde demanda. Agentes de IA não substituem o julgamento clínico; eles o ampliam. Não eliminam a necessidade de profissionais qualificados; reduzem sua carga cognitiva e administrativa. E não resolvem por si sós as desigualdades do sistema de saúde, mas podem ser ferramentas poderosas na direção de um cuidado mais equitativo, acessível e baseado em evidências.

O futuro imediato da IA na saúde será moldado pela capacidade do setor de estabelecer frameworks de governança robustos, datasets representativos da diversidade da população, protocolos de validação clínica rigorosos e culturas institucionais que promovam o uso crítico, e não a dependência acrítica, dessas ferramentas. Profissionais de saúde, desenvolvedores de tecnologia, reguladores e pacientes precisam participar conjuntamente desse processo de construção. Este minicurso é uma contribuição para a formação dos primeiros aqueles que, equipados com compreensão técnica e consciência ética, serão os protagonistas dessa transformação.

Referências

ALOMARI, L. M. et al. Safety and accuracy of ai in triaging patients in the emergency department. *International Journal of Emergency Medicine*, Springer, v. 18, n. 1, p. 243, 2025. Citado na página page.2020.

ANTHROPIC. *Model Context Protocol Introduction*. 2024. <https://modelcontextprotocol.io/introduction>. Acesso em: 10 maio 2026. Citado na página page.33.

Associação Nacional de Hospitais Privados (Anahp). *A Inteligência Artificial já está sendo usada na saúde: descubra como isso te beneficia*. 2025. Publicado com base nos dados apresentados no Showcase de IA nos Hospitais Brasileiros. Disponível em: <https://www.anahp.com.br/saude-da-saude/a-inteligencia-artificial-ja-esta-sendo-usada-na-saude-descubra-como-isso-te-beneficia/>. Citado na página page.22.

BASTIAN, H.; GLASZIOU, P.; CHALMERS, I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS medicine*, Public Library of Science San Francisco, USA, v. 7, n. 9, p. e1000326, 2010. Citado na página page.2323.

BHAVNANI, S. P.; NARULA, J.; SENGUPTA, P. P. Mobile technology and the digitization of healthcare. *European heart journal*, v. 37, n. 18, p. 1428, 2016. Citado na página page.2424.

COLLINS, F. S.; VARMUS, H. A new initiative on precision medicine. *New England journal of medicine*, Mass Medical Soc, v. 372, n. 9, p. 793–795, 2015. Citado na página page.2323.

Distrito; Associação Brasileira de Startups de Saúde e HealthTechs (ABSS). *HealthTech Recap 2024*. [S.l.], 2025. Relatório lançado em fevereiro de 2025 com dados do exercício de 2024. Disponível em: <https://materiais.distrito.me/report/healthtech-report-recap-2024>. Citado na página page.22.

ESTEVA, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, Nature Publishing Group UK London, v. 542, n. 7639, p. 115–118, 2017. Citado na página page.2121.

ESTEVA, A. et al. A guide to deep learning in healthcare. *Nature medicine*, Nature Publishing Group US New York, v. 25, n. 1, p. 24–29, 2019. Citado na página page.22.

GOODFELLOW, I. et al. *Deep learning*. [S.l.]: MIT press Cambridge, 2016. v. 1. Citado na página page.22.

GRABER, M. L. The incidence of diagnostic error in medicine. *BMJ quality & safety*, BMJ Publishing Group Ltd, v. 22, n. Suppl 2, p. ii21–ii27, 2013. Citado na página page.2222.

GULSHAN, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, American Medical Association, v. 316, n. 22, p. 2402–2410, 2016. Citado na página page.2121.

HABERLE, T. et al. The impact of nuance dax ambient listening ai documentation: a cohort study. *Journal of the American Medical Informatics Association*, Oxford University Press, v. 31, n. 4, p. 975–979, 2024. Citado na página page.2222.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3. ed. Burlington, MA: Morgan Kaufmann, 2011. ISBN 9780123814791. Disponível em: <https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques>. Citado na página page.33.

KRZYSZCZYK, P. et al. The growing role of precision and personalized medicine for cancer treatment. *Technology*, World Scientific, v. 6, n. 03n04, p. 79–100, 2018. Citado na página page.2323.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015. Disponível em: <https://www.nature.com/articles/nature14539>. Citado na página page.33.

LEWIS, P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, v. 33, p. 9459–9474, 2020. Disponível em: <https://arxiv.org/abs/2005.11401>. Citado na página page.33.

LINZER, M. et al. A cluster randomized trial of interventions to improve work conditions and clinician burnout in primary care: results from the healthy work place (hwp) study. *Journal of general internal medicine*, Springer, v. 30, n. 8, p. 1105–1111, 2015. Citado na página page.2323.

LIPPMANN, R. P. An introduction to computing with neural nets. *ACM SIGARCH Computer Architecture News*, ACM New York, NY, USA, v. 16, n. 1, p. 7–25, 1988. Citado na página page.33.

MARSHALL, I. J.; WALLACE, B. C. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic reviews*, Springer, v. 8, n. 1, p. 163, 2019. Citado na página page.2323.

MILLER, R. A.; JR, H. E. P.; MYERS, J. D. Internist-i, an experimental computer-based diagnostic consultant for general internal medicine. In: *Computer-assisted medical decision making*. [S.l.]: Springer, 1985. p. 139–158. Citado na página page.33.

MK, P.-O. The causal pathways linking health literacy to health outcomes. *Am J Health Behav.*, v. 31, n. 1, p. S19–S26, 2007. Citado na página page.2424.

MÜLLER, J. P. Architectures and applications of intelligent agents: A survey. *The Knowledge Engineering Review*, Cambridge University Press, v. 13, n. 4, p. 353–380, 1999. Citado na página page.55.

Núcleo de Informação e Coordenação do Ponto BR (NIC.br). *Pesquisa sobre o uso das tecnologias de informação e comunicação nos estabelecimentos de saúde brasileiros: TIC Saúde 2024*. São Paulo, 2024. Entrevistas realizadas entre fevereiro e agosto de 2024 com 2.057 gestores e 2.021 profissionais de saúde em todo o território nacional. Disponível em: <https://cetic.br/pt/tics/saude/2024/medicos/>. Citado na página page.22.

OBERMEYER, Z.; EMANUEL, E. J. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, v. 375, n. 13, p. 1216, 2016. Citado na página page.2121.

RAZAGHI, M. et al. Transforming clinical documentation with ambient artificial intelligence (ai) scribes: a narrative review of technology, impact, and implementation. *Cardiovascular Diagnosis and Therapy*, LWW, v. 16, n. 1, p. 11, 2026. Citado na página page.2222.

REZENDE, S. O. *Sistemas inteligentes: fundamentos e aplicações*. [S.l.]: Editora Manole Ltda, 2003. Citado na página page.22.

RUSSELL, S.; NORVIG, P.; INTELLIGENCE, A. A modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, v. 25, n. 27, p. 79–80, 1995. Citado 2 vezes nas páginas page.22 e page.55.

SHAN, R.; SARKAR, S.; MARTIN, S. S. Digital health technology and mobile devices for the management of diabetes mellitus: state of the art. *Diabetologia*, Springer, v. 62, n. 6, p. 877–887, 2019. Citado na página page.2424.

SHORTLIFFE, E. *Computer-based medical consultations: MYCIN*. [S.l.]: Elsevier, 2012. v. 2. Citado na página page.33.

SINGHAL, K. et al. Large language models encode clinical knowledge. *Nature*, Nature Publishing Group UK London, v. 620, n. 7972, p. 172–180, 2023. Citado na página page.2323.

SINSKY, C. et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Annals of internal medicine*, American College of Physicians, v. 165, n. 11, p. 753–760, 2016. Citado na página page.2222.

SOMASHEKHAR, S. P. et al. Watson for oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. *Annals of Oncology*, Elsevier, v. 29, n. 2, p. 418–423, 2018. Citado na página page.2121.

SUTTON, R. T. et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, Nature Publishing Group UK London, v. 3, n. 1, p. 17, 2020. Citado na página page.2121.

TOPOL, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, Nature Publishing Group US New York, v. 25, n. 1, p. 44–56, 2019. Citado 2 vezes nas páginas page.22 e page.2020.

VASWANI, A. et al. Attention is all you need. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. v. 30, p. 5998–6008. Disponível em: <https://papers.nips.cc/paper/7181-attention-is-all-you-need>. Citado 2 vezes nas páginas page.33 e page.88.

WOOLDRIDGE, M. *An introduction to multiagent systems*. [S.l.]: John wiley & sons, 2009. Citado na página page.22.

YUAN, K.-C. et al. The development an artificial intelligence algorithm for early sepsis diagnosis in the intensive care unit. *International journal of medical informatics*, Elsevier, v. 141, p. 104176, 2020. Citado na página page.2121.

