Capítulo 3

Construção de Agente de Conversação com Aplicação na Saúde: Do Problema Real à Solução Inteligente

Amanda Morais Almeida, Diêgo Farias de Freitas, Paulo Eduardo Ambrósio

Abstract

This chapter addresses the development of intelligent conversational agents as a solution to structural challenges and opportunities in the healthcare sector. The proposal is based on the need for Artificial Intelligence (AI) adoption to automate processes, minimize errors, and reduce the cognitive load of professionals, utilizing an architecture grounded in Large Language Models (LLMs) integrated with Retrieval-Augmented Generation (RAG). This study describes essential technical foundations, including the role of vector embeddings, context management via chunking, and the orchestration of clinical workflows through the LangChain framework. To ensure patient safety and mitigate hallucinations, it discusses the implementation of technical supervision layers called Guardrails and the use of multi-agent systems for decision auditing. The text concludes by analyzing implementation challenges in heterogeneous infrastructures, the necessity of compliance with regulatory frameworks, and perspectives for digital health that keeps the human in the loop

Resumo

Este capítulo aborda o desenvolvimento de agentes conversacionais inteligentes como solução para desafios e oportunidades estruturais no setor de saúde. A proposta parte da necessidade de adesão à Inteligência Artificial (IA) para a automatização de processos, minimização de erros e redução da carga cognitiva de profissionais, utilizando uma arquitetura baseada em Modelos de Linguagem de Grande Escala (LLMs) integrados à Geração Aumentada por Recuperação (RAG). A presente investigação descreve os fundamentos técnicos essenciais, incluindo o papel dos embeddings vetoriais, a gestão de contexto via chunking e a orquestração de fluxos clínicos através do framework LangChain. Para garantir a segurança do paciente e mitigar alucinações, é discutida a implementação de camadas de supervisão técnica denominadas Guardrails e o uso de sistemas multiagentes para a auditoria de decisões. O texto conclui analisando os desafios de implementação em infraestruturas heterogêneas, a necessidade de conformidade com marcos regulatórios e as perspectivas para uma saúde digital que mantenha o ser humano no centro do processo (human-in-the-loop).

3.1. Agentes de IA na Transformação Digital da Saúde

A Inteligência Artificial (IA) tem desempenhado um papel transformador no setor de saúde, embora grande parte de suas aplicações atuais ainda opere de forma restrita a tarefas isoladas, enfrentando barreiras como a complexidade dos dados clínicos e a presença de vieses algorítmicos. Essa limitação tecnológica ocorre em um momento de crise estrutural no setor que, de acordo com Karunanayake (2025), projeta-se que haverá um déficit global de 18 milhões de profissionais de saúde até 2030. Esse dado evidencia a possibilidade do significativo impacto no acesso ao cuidado, especialmente em regiões de baixa renda. Diante desse cenário, a integração de tecnologias digitais deixa de ser apenas uma inovação para se tornar uma estratégia fundamental dos sistemas médicos modernos, visando a otimização de fluxos de trabalho e a democratização de soluções inteligentes.

No âmbito prático dessa transformação, a eficiência da triagem em departamentos de emergência é frequentemente comprometida pela natureza dinâmica do ambiente clínico. Segundo Suamchaiyaphum et al. (2024), a acurácia dos enfermeiros é diretamente influenciada por fatores como o volume de pacientes e a carga horária de trabalho, o que pode gerar um desalinhamento entre a necessidade clínica e a decisão tomada. Essa lacuna de conhecimento e a variabilidade nos resultados de triagem evidenciam a urgência de sistemas de apoio que reduzam a carga cognitiva e garantam a aderência aos protocolos, independentemente das pressões ambientais.

Adicionalmente, a aplicação da triagem na Atenção Primária à Saúde (APS) revela desafios distintos. Conforme apontado por Park et al. (2025), embora a triagem conduzida por enfermeiros apresente níveis aceitáveis de concordância clínica, há uma tendência persistente ao over-referral que é descrito como encaminhamento excessivo para serviços de emergência como estratégia de segurança. Nesse cenário, a transição para

sistemas baseados em IA e *machine learning* surge não apenas como uma inovação, mas como uma necessidade para aprimorar a priorização clínica. O suporte de ferramentas digitais e algoritmos validados permite que o profissional de saúde atue de forma mais precisa na seleção do local de cuidado, otimizando os recursos do sistema e garantindo que o acesso do paciente seja pautado por evidências estruturadas, reduzindo a subjetividade inerente ao atendimento remoto ou síncrono.

Para mitigar tais limitações, a transição para sistemas de maior confiabilidade e com arquitetura multicamadas marca o início da chamada era agêntica. Conforme detalha Karunanayake (2025), este momento é caracterizado por sistemas de agentes de IA com funcionalidade autônoma, raciocínio avançado e interações dinâmicas. Diferente das aplicações tradicionais, essa tecnologia incorpora arquiteturas adaptáveis que aumentam a autonomia na tomada de decisão. No contexto da triagem, isso significa que o sistema utiliza um raciocínio probabilístico para atualizar previsões continuamente com base em novos dados, funcionando como uma solução capaz de mitigar a escassez de profissionais ao automatizar tarefas administrativas e aprimorar a eficiência do fluxo de trabalho na porta de entrada dos serviços de saúde.

Contudo, essa transformação digital não se resume à automação de processos, mas à criação de um ecossistema onde a inteligência artificial atua como um extensor das capacidades humanas. Sob a perspectiva da Modelagem Computacional, o desafio reside na representação fiel dos protocolos clínicos em modelos de decisão que sejam, simultaneamente, flexíveis para a linguagem natural e rígidos para a segurança do paciente. Assim, a era agêntica na saúde pressupõe uma infraestrutura que suporte o aprendizado contínuo e a explicabilidade, transformando o vasto volume de dados clínicos em estratégia para otimização de fluxos operacionais, maior segurança na prestação de serviços e no desenvolvimento de produtos para os pacientes.

3.2. Fundamentos Técnicos para Construção de Agentes

Este tópico dedica-se à construção da base teórica necessária para viabilizar o desenvolvimento do agente, permitindo uma compreensão aprofundada de como cada componente técnico influencia diretamente no desempenho e na confiabilidade do resultado final. A investigação percorre as camadas fundamentais do sistema com o objetivo de demonstrar a atuação de cada tecnologia e os estudos existentes que validam sua eficácia em cenários de alta criticidade. Ao detalhar as definições conceituais e suas finalidades específicas, o texto estabelece as metodologias adotadas e as correlaciona com as alternativas e paradigmas de automação vigentes na literatura.

A fundamentação parte de uma análise dos componentes arquiteturais que permitem a transformação de uma necessidade real em uma solução inteligente e tecnicamente segura. Para isso, o capítulo aprofunda-se no Processamento de Linguagem Natural (PLN) e no papel dos *Large Language Models* (LLM) como núcleos de processamento cognitivo. Em seguida, são explorados os *embeddings* como técnica de representação vetorial e as estratégias de *chunking* e janelas de memória, elementos

essenciais para a gestão de contexto que compõe a arquitetura de *Retrieval-Augmented Generation* (RAG), traduzida como Geração Aumentada por Recuperação. Esta estrutura é refinada pelas técnicas de engenharia de *prompt*, que otimizam a interação. Por fim, abordam-se os *guardrails*, voltados à garantia da segurança sistêmica, e a orquestração via LangChain, que viabiliza a integração eficiente de todas as camadas do agente.

A compreensão dos agentes inteligentes no setor saúde exige a análise das tecnologias que permitem a transição de sistemas estáticos para modelos dinâmicos de interação. A arquitetura desses sistemas baseia-se em avanços significativos no Processamento de Linguagem Natural (PLN) e na Ciência de Dados, permitindo que diferentes classes de agentes desempenhem funções específicas no fluxo hospitalar. A Tabela 1 sintetiza essa diversidade tecnológica, categorizando os agentes de acordo com suas aplicações, categorias de saúde e bases computacionais [Karunanayake 2025].

Tabela 3.1. Categorização de tipos de agentes de IA na saúde [Karunanayake 2025].

Agentes de IA	Principais Aplicações	Categorias de Saúde	Principais Usuários	Principais Tecnologias de IA
Agentes baseados em imagem	Diagnóstico de doenças, detecção precoce, geração de relatórios.	Diagnóstico, suporte à decisão clínica.	Radiologistas, Médicos.	Visão computacional (CNNs, ViTs), MLLMs para integração imagem-texto.
Agentes de análise preditiva	Predição de risco, previsão de progressão de doenças, resultados dos pacientes.	Suporte à Decisão Clínica, Tratamento e Cuidado, Descoberta de Fármacos.	Médicos, Equipes de Cuidado.	Modelagem Preditiva (ML supervisionado, ensemble, séries temporais).
Agentes conversacio-Nais	Verificação de sintomas, triagem de pacientes, consultas virtuais.	Engajamento e Monitoramento do Paciente.	Pacientes, Clínicos Gerais.	PLN, Sistemas de Diálogo, LLMs Pré-treinados.
Agentes de PLN (NLP)	Processamento de notas clínicas, sumarização de EHRs, extração de insights.	Operações e Administração, Suporte à Decisão Clínica.	Codificadores Médicos, Analistas.	PLN, LLMs Pré-treinados.
Agentes baseados em regras	Seguimento de diretrizes clínicas, alerta para interações medicamentosas.	Suporte à Decisão Clínica.	Médicos, Farmacêuticos.	Raciocínio Baseado em Regras, lógica, grafos de conhecimento.

Agentes híbridos	Combinação de imagem, texto, vídeo e análise preditiva para decisões.	Suporte à Decisão Clínica, Diagnóstico, Cirurgia Robótica.	Médicos, Radiologistas , Cirurgiões.	Aprendizado Multimodal.
Agentes de ML	Classificação de doenças, detecção de anomalias, planejamento de tratamento.	Diagnóstico, Tratamento, Descoberta de Fármacos.	Cientistas de Dados, Médicos.	Algoritmos de ML/DL, Aprendizado por Reforço (RL).
Agentes de sistemas especialistas	Emulação de expertise clínica para diagnóstico e planejamento.	Suporte à Decisão Clínica, Cirurgia Robótica.	Especialistas, Pesquisadores , Cirurgiões.	Sistemas baseados em conhecimento e regras.
Agentes recomenda-Dores	Sugestão de testes diagnósticos, tratamentos personalizados.	Tratamento e Cuidado, Suporte à Decisão Clínica.	Médicos, Equipes de Cuidado.	Filtragem colaborativa, sistemas de recomendação, RL.

A análise da Tabela 3.1 revela que o desenvolvimento de agentes conversacionais e sistemas híbridos para triagem clínica depende fundamentalmente de LLMs pré-treinados e de arquiteturas multimodais. A orquestração desses componentes visa transformar a entrada de dados não estruturados em decisões clínicas seguras e rastreáveis. Para compreender como essa conversão de linguagem bruta em decisão estruturada ocorre, é imperativo analisar o Processamento de Linguagem Natural como a camada primária de interface e inteligência do sistema.

3.2.1. Processamento de Linguagem Natural (PLN)

O Processamento de Linguagem Natural (PLN) é a disciplina da Inteligência Artificial que permite que máquinas realizem os processos de leitura e interpretação da comunicação humana. Segundo Jurafsky e Martin (2026), o principal desafio do PLN não é apenas identificar palavras, mas construir representações contextuais de significado. Para um sistema de saúde, isso significa entender que a palavra “aguda” em uma triagem possui um peso semântico de urgência que termos administrativos não possuem.

Para compreender a fundamentação dos agentes inteligentes atuais, é necessário observar a trajetória tecnológica que moldou o campo. Esta evolução, conforme discutido por Zaniboni (2025) e sintetizada por Jeevan (2023), reflete a transição de métodos estatísticos para modelos de entendimento profundo. Essa progressão histórica e as principais características de cada período estão detalhadas na Tabela 3.2.

Tabela 3.2 – Eras da evolução do Processamento de Linguagem Natural [Jeevan 2023].

Era	Período	Abordagem Principal	Características e Limitações
Baseada em Regras	1960 — 1980	Lógica simbólica (*se-então*)	Operava com caminhos rígidos criados por especialistas. Falhava em entender nuances, gírias ou contextos informais.
Estatística	1980 — 2000	Frequência e Probabilidade (ex: TF-IDF)	Identificava termos importantes pela frequência, mas ignorava a ordem das palavras e o significado real das frases.
Redes Neurais e *Embeddings*	2000 — 2017	Vetores Densos e RNNs	Palavras similares passaram a ser vizinhas em um espaço matemático. Contudo, a leitura sequencial dificultava o contexto em frases longas.
Era Agêntica	2017 — Presente	Transformers e Atenção	Processamento paralelo e global. O modelo consegue correlacionar informações distantes em um mesmo texto.

A inovação que sustenta os agentes modernos é a arquitetura Transformer, que introduziu o conceito de autoatenção para resolver o problema da ambiguidade [Vaswani et al. 2017]. Na prática, a atenção funciona como um filtro seletivo que permite ao modelo identificar quais palavras em uma frase são mais importantes para definir o sentido de um termo específico. À medida que um *token*, uma palavra ou parte dela, percorre esse fluxo, cada camada do Transformer adiciona informações contextuais a ele, sem apagar a informação original.

Jurafsky e Martin (2026) explicam que essa arquitetura opera através de um fluxo residual, onde cada camada do sistema adiciona novas camadas de entendimento sobre a palavra original sem descartar o que foi processado anteriormente. Ou seja, o modelo consegue realizar verificações para trás e para frente simultaneamente, permitindo que o significado de um termo seja enriquecido por palavras que apareceram muito antes no texto. Isso permite que um agente identifique, por exemplo, que o termo "pressão" em uma conversa clínica refere-se à medida arterial do paciente e não a uma força física ou emocional, baseando-se nas palavras vizinhas como "hipertensão" ou "aparelho".

O componente que permite essa possibilidade é a autoatenção ou *self-attention*. Didaticamente, a atenção funciona como um filtro de relevância. Quando o agente

processa a frase "O paciente apresenta dor intensa no peito", o mecanismo de atenção faz com que a palavra "intensa" indique com maior relevância para o modelo quando ele analisa "dor", enquanto palavras como "no" ou "o" são ignoradas. Matematicamente, isso é feito através de três vetores: *Query* (Consulta), *Key* (Chave) e *Value* (Valor). No exemplo citado, o vetor *Query* pergunta "O que eu estou procurando?", o *Key* responde "O que eu tenho de informação?" e o *Value* entrega o conteúdo relevante. Essa interação resulta em Mapas de Atenção que são visualizações que mostram as conexões que o modelo faz entre as palavras.

Para ilustrar esse funcionamento, a Figura 3.1 apresenta o comportamento das *attention heads* ou cabeças de atenção, ao processarem a palavra *it*, traduzida como ele ou ela, em uma sentença. Nota-se que, enquanto uma cabeça foca no sujeito *the animal*, outra foca no adjetivo *tired*. Em termos de modelagem, isso significa que a representação final da palavra *it* funde ou absorve parte das características de ambos os termos, permitindo que o modelo compreenda o contexto completo da frase.

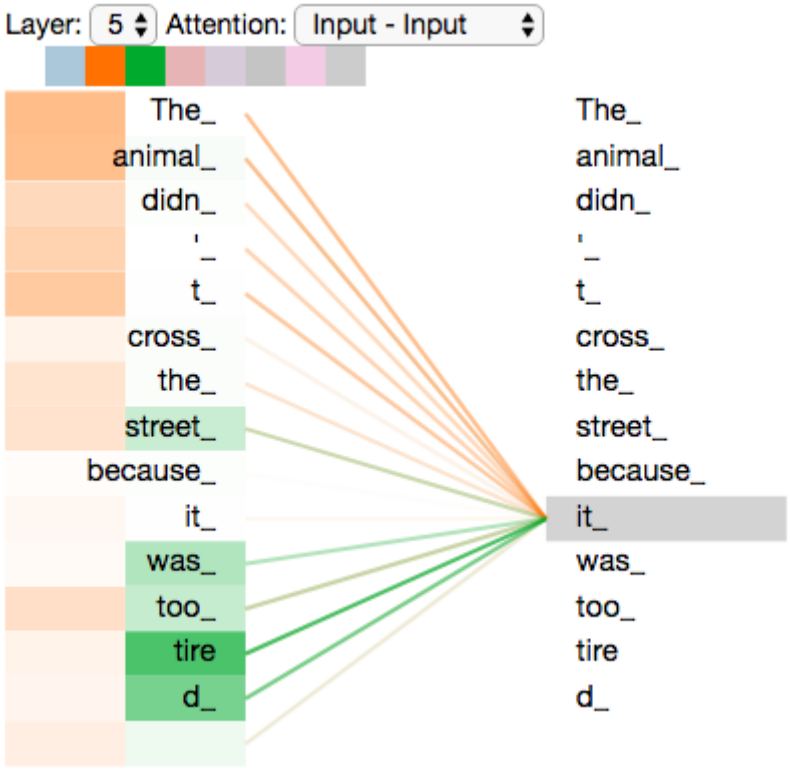

Figura 3.1 – Visualização das *attention heads* focando em diferentes dependências semânticas [Alammar 2018].

O desenvolvimento de um agente inteligente depende dessa capacidade de capturar o conhecimento de mundo através do pré-treinamento generativo [Radford et al. 2018]. Esse processo permite que o modelo aprenda a estrutura da linguagem em grandes acervos de livros e documentos antes de ser ajustado para a tarefa específica. Para a

construção do agente, essa base teórica garante que o sistema não apenas repita frases programadas, mas que consiga realizar transformações de entrada para lidar com diferentes tipos de perguntas e interações, mantendo a coerência e a lógica durante todo o evento.

Na área da saúde, PLN exerce um papel estratégico, uma vez que viabiliza o tratamento da desestruturação típica de prontuários e relatos médicos [Nascimento 2024]. O sistema consegue realizar a extração de entidades nomeadas, identificando automaticamente sintomas e dosagens de medicamentos em meio a textos longos, o que agiliza o fluxo de trabalho hospitalar e reduz a carga cognitiva da equipe assistencial. Além disso, a análise de sentimento permite ao agente capturar o estado emocional ou o nível de desconforto relatado pelo paciente, oferecendo uma camada de humanização e precisão diagnóstica que sistemas puramente estatísticos não conseguiriam atingir.

A integração dessas camadas tecnológicas resulta na criação de sistemas que ao processar palavras, interpretam a gravidade e a urgência contidas na linguagem natural. Esta base teórica é o que permite que os *Large Language Models* (LLMs) compreendam a fundo as necessidades dos pacientes, elevando a capacidade de resposta do agente a níveis de precisão comparáveis ao entendimento humano em tarefas específicas de suporte à decisão.

3.2.2. *Large Language Models* (LLM)

Os Modelos de Linguagem de Grande Escala, amplamente conhecidos pela sigla LLM, constituem o estágio mais avançado na evolução do Processamento de Linguagem Natural. Em termos estruturais, um LLM é um modelo baseado na arquitetura Transformer que foi treinado em uma escala sem precedentes de dados e parâmetros. Segundo Brown et al. (2020), a principal característica que define esses sistemas é o surgimento de habilidades cognitivas complexas conforme o modelo é escalado para centenas de bilhões de parâmetros, como observado no GPT-3. Para a literatura, esse fenômeno de escala é quase comparável a uma lei da física, onde capacidades imprevistas de raciocínio e síntese emergem de forma previsível à medida que o volume de dados e o poder computacional aumentam [Tamkin et al. 2021].

O diferencial fundamental entre um LLM e as ferramentas de PLN tradicionais reside na sua natureza generalista. Enquanto modelos antigos eram treinados para tarefas específicas, como classificar sentimentos ou traduzir frases, os LLMs são definidos como "aprendizes de poucos exemplos" ou *few-shot learners*. Isso significa que eles possuem a capacidade intrínseca de compreender e executar novas tarefas recebendo apenas algumas demonstrações textuais dentro do contexto da conversa, sem a necessidade de atualizar seus pesos matemáticos ou passar por um novo treinamento [Brown et al. 2020].

Para ilustrar essa versatilidade, a Figura 3.2 apresenta as três configurações principais de uso desses modelos. No modo *zero-shot*, o modelo recebe apenas uma descrição da tarefa, no *one-shot*, é fornecido um único exemplo de demonstração e no *few-shot*, o modelo é condicionado com alguns exemplos antes da execução final. Nota-se que, ao contrário do ajuste fino tradicional conhecido como *fine-tuning*, essas abordagens não realizam atualizações de gradiente, operando puramente por meio da interação textual.

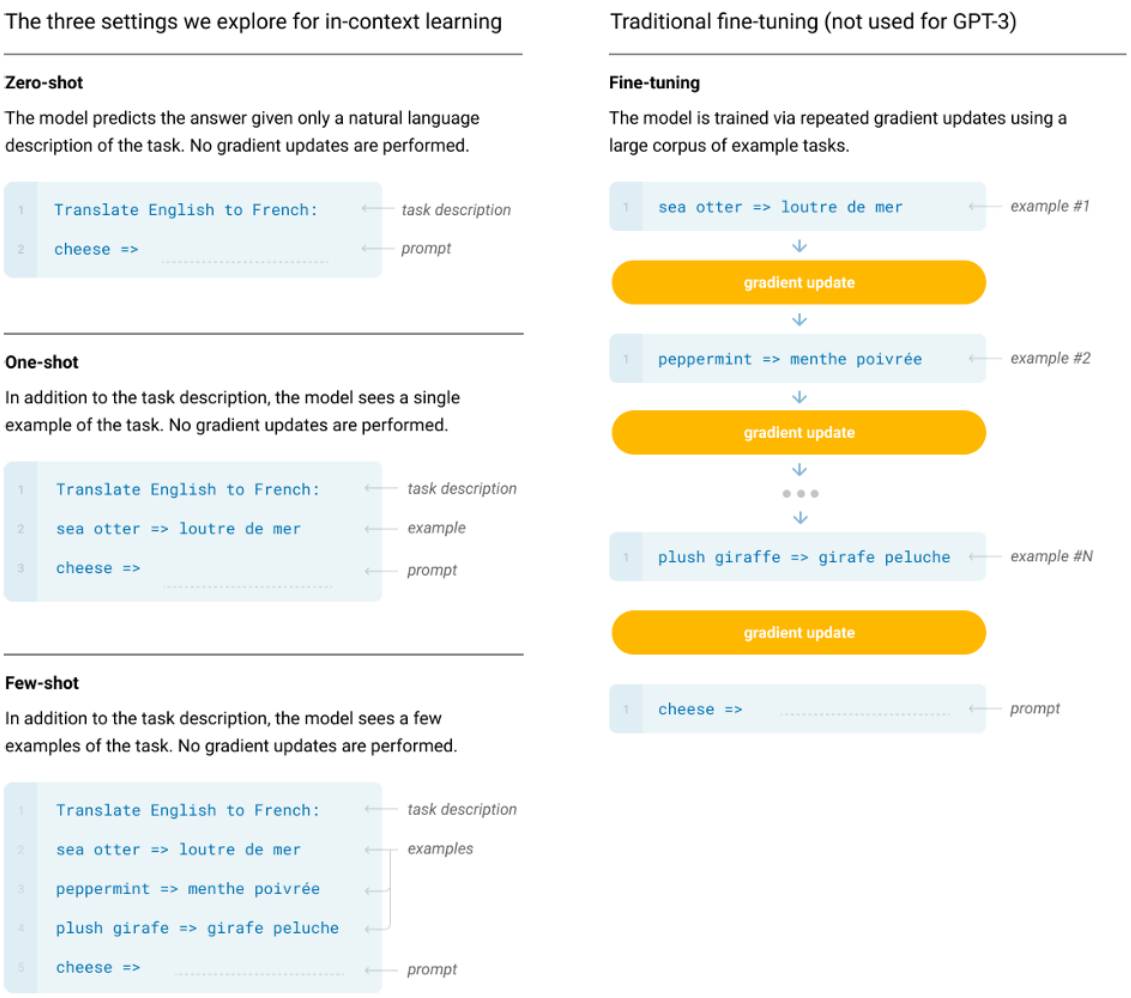

Figura 3.2 – Comparativo entre as abordagens de aprendizado de máquina [Brown et al. 2020]

O funcionamento desses modelos baseia-se no pré-treinamento generativo, um processo em que o sistema consome trilhões de *tokens* extraídos de livros, artigos científicos, códigos de programação e conversas da internet [Touvron et al. 2023]. Esse estágio inicial permite que o modelo capture um vasto conhecimento de mundo e aprenda a estrutura profunda da linguagem. No entanto, o treinamento bruto não é suficiente para garantir um comportamento útil ou seguro. Conforme discutido por Tamkin et al. (2021), é necessário realizar um processo denominado de alinhamento, que ajusta os objetivos

estatísticos do modelo aos valores e intenções humanas, garantindo que o agente inteligente responda de forma útil e ética.

Didaticamente, o uso de um LLM para a construção de agentes inteligentes ocorre por meio do *design* de comandos, ou engenharia de *prompts*. Diferente da programação tradicional, onde o desenvolvedor define regras lógicas rígidas, o sistema é guiado por linguagem natural. O LLM atua como o núcleo de processamento do agente, sendo capaz de manter a coerência em diálogos longos, resumir informações complexas e gerar respostas estruturadas. Segundo Touvron et al. (2023), modelos menores e mais eficientes, como a série LLaMA, demonstraram que o refinamento dos dados de treinamento pode ser tão impactante quanto o número de parâmetros, permitindo o uso dessas tecnologias em infraestruturas locais de pesquisa.

A diversidade de LLMs disponíveis atualmente pode ser classificada primordialmente entre modelos proprietários e modelos de código aberto denominados *open-source*. Modelos proprietários, como a série GPT da OpenAI, geralmente oferecem as maiores escalas de parâmetros e performance de estado da arte, mas operam sob acesso restrito via API, sigla para Interface de Programação de Aplicações, o que limita a transparência sobre os dados de treinamento e o controle sobre a privacidade das informações [Brown et al. 2020]. Em contrapartida, coleções de modelos abertos, como o LLaMA da Meta AI, democratizaram o acesso à tecnologia ao permitir que modelos competitivos, variando de 7B a 65B de parâmetros, sejam executados em infraestruturas de pesquisa locais, garantindo maior controle sobre o fluxo de dados e a auditoria do sistema [Touvron et al. 2023].

A escolha do modelo ideal para uma aplicação específica deve considerar o equilíbrio entre o desempenho desejado e o orçamento de inferência. Conforme demonstrado por Touvron et al. (2023), o desempenho não é determinado exclusivamente pelo tamanho do modelo. Modelos menores treinados com um volume massivo de dados podem superar modelos gigantes, como o GPT-3 de 175B, em diversos *benchmarks*. Para o desenvolvimento de agentes inteligentes, modelos mais compactos e eficientes são frequentemente preferíveis, pois reduzem a latência na resposta e os custos computacionais sem comprometer a precisão necessária para tarefas de raciocínio e síntese.

Além da escala, a seleção deve ser guiada pela capacidade de aprendizado em contexto e pelo alinhamento com a tarefa pretendida [Tamkin et al. 2021]. Em cenários de alta sensibilidade, como a saúde, deve-se priorizar modelos que apresentem menores taxas de alucinação e maior resiliência a vieses, avaliados por métricas de veracidade como o *TruthfulQA*. Assim, a modelagem de um agente inteligente clínico exige uma avaliação criteriosa se o modelo escolhido possui a superfície de capacidade necessária para processar a linguagem técnica biomédica mantendo a conformidade com os requisitos éticos de segurança e privacidade [Weidinger et al. 2021].

As contribuições dos LLMs para a ciência moderna são significativas, especialmente na democratização do acesso a ferramentas de inteligência avançada. Eles permitem a tradução de nuances culturais que sistemas antigos ignoravam e a síntese de grandes volumes de literatura técnica em segundos. Exemplos práticos incluem assistentes virtuais capazes de depurar códigos de programação complexos, sistemas de escrita colaborativa e ferramentas de busca que, em vez de apenas listar *links*, constroem respostas fundamentadas e contextualizadas.

No cenário da saúde, a transição para os LLMs potencializa as técnicas de PLN citadas anteriormente, permitindo que a interpretação de prontuários ultrapasse a mera identificação de palavras-chave para alcançar uma compreensão semântica profunda dos relatos clínicos. A capacidade generativa permite que o agente não apenas reconheça dados desestruturados, mas sintetize históricos complexos, infira relações de causalidade entre sintomas e organize informações biomédicas em formatos padronizados de forma automatizada. Essa evolução transforma o processamento de texto em uma ferramenta ativa de suporte à decisão, agilizando fluxos assistenciais em ambientes de alta criticidade [Brown et al. 2020]. A arquitetura desse processamento clínico, conforme detalhada na Figura 3.3, ilustra justamente o fluxo de transformação do relato bruto do paciente em uma saída técnica estruturada.

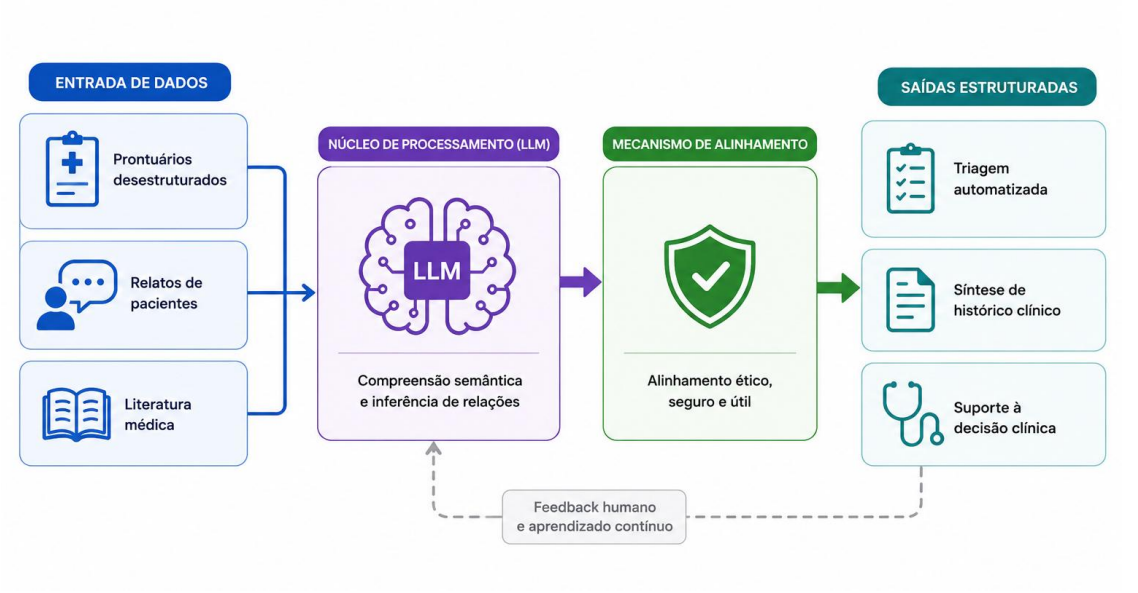

Figura 3.3 – Arquitetura de processamento clínico via LLM

Entretanto, as implicações do uso de LLMs na saúde exigem uma governança rigorosa. Weidinger et al. (2021) alertam para o risco de desinformação médica, onde o modelo, por ser puramente estatístico, pode gerar orientações clinicamente perigosas que parecem factuais mas são, na verdade, alucinações. Como os modelos aprendem padrões de associação e não a verdade científica intrínseca, eles podem falhar em distinguir entre

conselhos médicos baseados em evidências e desinformação presente em seus dados de treinamento [Weidinger et al. 2021].

Outro ponto de atenção reside na reprodução de estereótipos sociais nocivos. LLMs treinados em vastos conjuntos de dados da internet podem espelhar e até amplificar preconceitos de raça, gênero ou religião [Touvron et al. 2023]. Em um ambiente de diagnóstico médico, essa propagação de vieses pode levar a um atendimento desigual para grupos minoritários, conforme alertado na taxonomia de riscos éticos desses modelos [Weidinger et al. 2021]. A literatura enfatiza que a representação injusta de identidades sociais pode criar danos materiais diretos se o sistema for utilizado para alocação de recursos hospitalares ou definição de prioridades clínicas.

Além dos riscos de conteúdo, a segurança da informação é uma preocupação latente. Os LLMs podem, inadvertidamente, memorizar e vazar informações sensíveis ou dados privados que estavam presentes em seus acervos de treinamento [Weidinger et al. 2021]. No contexto da saúde, onde a privacidade dos dados do paciente é um requisito ético e legal rigoroso, a implementação desses modelos requer camadas adicionais de anonimização e técnicas de treinamento que garantam a integridade contra vazamentos acidentais.

A interação humana com esses modelos também apresenta desafios psicológicos. O fenômeno da atribuição de características humanas pode levar pacientes a superdimensionarem as competências do agente inteligente por ele falar de maneira fluida e humanizada [Weidinger et al. 2021]. Essa confiança indevida pode fazer com que usuários ignorem a necessidade de supervisão médica ou aceitem diagnósticos automatizados sem o devido ceticismo clínico, o que reforça a importância de manter sempre um profissional humano no ciclo de decisão.

Para estruturar esses riscos, a Tabela 3.3 resume a taxonomia de possíveis danos associados aos modelos de linguagem de grande escala. Essa classificação é essencial para orientar o desenvolvimento de estratégias de mitigação eficazes.

Tabela 3.3 – Taxonomia de riscos éticos e sociais associados aos LLMs [Weidinger et al. 2021]

Área de Risco	Mecanismo de Origem	Tipos de Dano
I. Discriminação, Exclusão e Toxicidade	O modelo reflete com precisão padrões de fala injustos, tóxicos ou opressores presentes nos dados de treino.	Ofensa, danos materiais (locacionais) e tratamento injusto ou representação estereotipada de grupos marginalizados.
II. Riscos de Informação (Information Hazards)	O modelo prevê enunciados que constituem informações privadas ou críticas para a segurança presentes ou inferidas dos dados de treino.	Violações de privacidade de dados sensíveis e riscos diretos à segurança do indivíduo.
III. Danos por Desinformação	O modelo atribui altas probabilidades a informações falsas, enganosas, sem sentido ou de baixa qualidade.	Engano, danos materiais e ações antiéticas por humanos que aceitam a previsão do modelo como um facto, além da erosão da confiança social.
IV. Usos Maliciosos	Utilização intencional do modelo por atores humanos com o objetivo explícito de causar dano.	Enfraquecimento do discurso público, fraudes, campanhas de desinformação personalizadas e produção de código informático malicioso.
V. Danos na Interação Humano-Computador	Riscos decorrentes da interação direta via diálogo, onde o utilizador interage com o sistema como se fosse um agente humano.	Uso inseguro devido ao erro de julgamento do utilizador, vulnerabilidades psicológicas e perpetuação de estereótipos via design do produto (ex: assistentes femininas).
VI. Automação, Acesso e Danos Ambientais	Utilização de LLMs em aplicações que beneficiam desproporcionalmente alguns grupos em detrimento de outros.	Aumento das desigualdades sociais, perda de empregos de alta qualidade, desigualdade no acesso à tecnologia e custos ambientais elevados.

Para mitigar esses riscos, a literatura sugere a aplicação de protocolos de inovação responsável e testes rigorosos de segurança, conhecidos como *red-teaming* [Tamkin et al. 2021]. Não basta que o modelo seja tecnicamente avançado, ele deve ser alinhado e constantemente verificado contra falhas éticas. A maturidade técnica dos LLMs oferece

mecanismos necessários para interpretar a linguagem natural com precisão, entretanto é a governança desses sistemas que garantirá sua utilidade clínica.

3.2.3. *Embeddings* para Representação Vetorial

Os *embeddings* constituem o mecanismo fundamental que permite aos LLMs compreenderem a semântica de um texto. Eles são representações matemáticas de palavras, frases ou documentos inteiros em um espaço vetorial de alta dimensionalidade. Diferentemente de representações simbólicas tradicionais, os *embeddings* convertem texto em vetores numéricos onde a proximidade geométrica entre dois pontos reflete a similaridade semântica dos termos que eles representam [Jeevan 2023].

Na prática, o processo envolve submeter o texto a um modelo de linguagem capaz de gerar *embeddings*, como o BERT ou modelos de *embedding* da OpenAI, que produzem vetores numéricos correspondentes. Esses vetores podem ser armazenados em bancos de dados vetoriais, permitindo consultas eficientes baseadas em similaridade.

A Figura 3.4 ilustra como diferentes conceitos médicos são mapeados em um espaço tridimensional, mostrando a proximidade entre termos semanticamente relacionados.

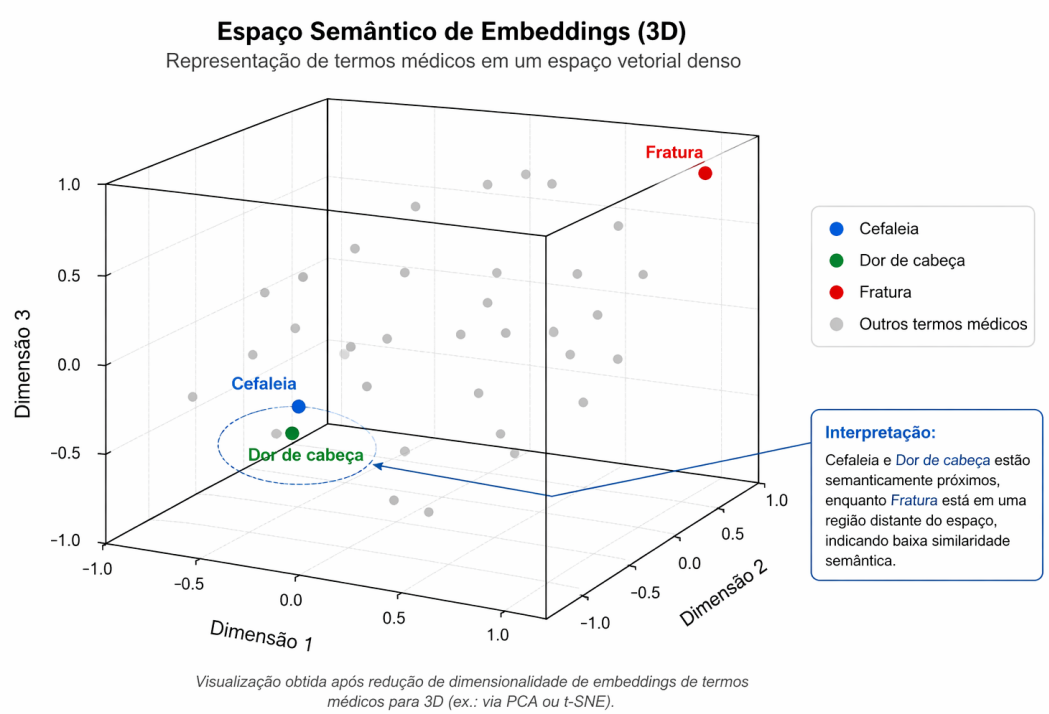

Figura 3.4 – Visualização conceitual de *embeddings* médicos em espaço vetorial

A representação evidencia que termos semanticamente semelhantes ocupam regiões próximas no espaço vetorial, enquanto conceitos distintos apresentam maior

separação. Existem diversos tipos de modelos de embedding, desde os mais simples e estáticos, como o modelo Word2Vec, até os mais modernos, profundos e contextuais baseados em arquiteturas Transformers. Para a área da saúde, é comum o uso de modelos pré-treinados em grandes *corpora* de literatura biomédica, como o BioBERT, que capturam com maior precisão as nuances do vocabulário técnico.

Diferente de modelos estáticos, onde uma palavra possui um único vetor invariável, os *embeddings* modernos são contextuais. Isso permite que o sistema diferencie o sentido de termos polissêmicos com base nas palavras vizinhas, o que é importante para reduzir ambiguidades em diagnósticos [Tamkin et al. 2021]. A principal implicação do uso de *embeddings* na saúde reside, portanto, na superação da busca por palavras-chave exatas. Em um sistema de triagem, isso permite que o agente identifique que os termos "cefaléia intensa" e "dor de cabeça aguda" ocupam regiões próximas no espaço vetorial, ainda que não compartilhem a mesma grafia. Segundo Lewis et al. (2020), essa representação é a base para a recuperação de informações em larga escala, permitindo que o modelo de linguagem processe conceitos clínicos complexos de forma estruturada e matematicamente comparável.

3.2.4. *Chunking* e Janelas de Memória como Gestão de Contexto

A gestão de contexto refere-se às técnicas utilizadas para lidar com o chamado limite de tokens ou limite de contexto dos LLMs. Este limite é a quantidade máxima de informação que o modelo consegue processar de uma só vez. As duas estratégias principais são o *chunking* e as janelas de memória. O *chunking* é o processo de segmentar documentos extensos em fragmentos menores e logicamente coerentes [Dong et al. 2024]. Já as janelas de memória constituem técnicas de filtragem e retenção que garantem que o agente selecione e recupere as partes cruciais de uma interação em uma conversa de longa duração, evitando a saturação do limite de *tokens* e a perda de informações relevantes [Akheel 2025].

O *chunking* é aplicado antes de converter o texto em *embeddings*, dividindo os documentos médicos, como protocolos hospitalares, em blocos de tamanho fixo ou baseados em parágrafos. Um fator crítico nesta etapa é a sobreposição entre os blocos, chamada de *overlap*, que garante que a transição entre um fragmento e outro não resulte em perda de significado. As janelas de memória são configuradas no desenvolvimento do agente, definindo, por exemplo, que apenas as últimas dez interações ou um resumo do diálogo sejam mantidos no contexto para a próxima resposta.

O *chunking* pode ser simples com tamanho fixo de *tokens* ou inteligente, baseado na estrutura do texto, como títulos e parágrafos. Uma das estratégias mais eficazes é o Recursive Character Text Splitting, que tenta manter unidades maiores intactas, como parágrafos, recorrendo a divisões menores, sendo sentenças ou palavras apenas quando necessário para respeitar o limite de tamanho [Dong et al. 2024]. Para as janelas de

memória, existem tipos como *Buffer Memory* que guarda todo o histórico até o limite e *Summary Memory* a qual resume o histórico periodicamente [Akheel 2025].

A Figura 3.5 demonstra o processo de segmentação de texto (*chunking*) com sobreposição (*overlap*), evidenciando como fragmentos consecutivos compartilham partes do conteúdo para preservar o contexto semântico.

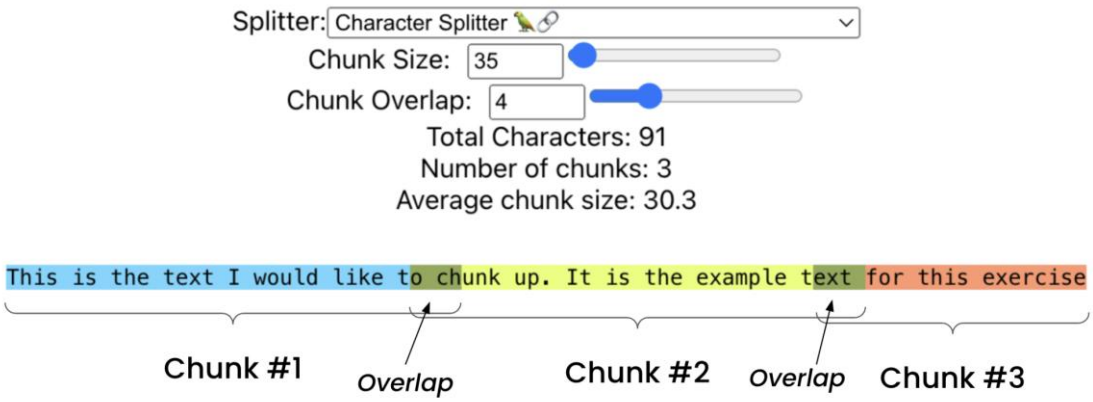

Figura 3.5 – Visualização do mecanismo de fragmentação com sobreposição (*overlap*) [Kamradt 2024]

Em um ambiente hospitalar, o *chunking* deve ser estratégico para não romper a continuidade de informações importantes como por exemplo, os sintomas de um paciente e sua respectiva duração devem permanecer no mesmo fragmento para que o modelo mantenha a coerência diagnóstica. A ausência de *overlap* pode resultar na fragmentação de informações críticas, comprometendo a análise semântica. Da mesma forma, uma janela de memória mal configurada pode fazer o agente não contemplar as alergias relatadas pelo paciente no início da conversa, gerando riscos clínicos graves. Segundo Akheel (2025), a escolha do tamanho desses fragmentos e a sobreposição entre eles são fundamentais para evitar a perda de contexto semântico durante a inferência.

3.2.5. *Retrieval-Augmented Generation* (RAG)

A Geração Aumentada por Recuperação, amplamente conhecida pela sigla RAG (*Retrieval-Augmented Generation*), é uma arquitetura de sistemas que combina as capacidades de recuperação de informação com a capacidade gerativa dos LLMs. O objetivo primordial é fornecer ao modelo informações externas, específicas e atualizadas, permitindo a geração de respostas tecnicamente precisas e fundamentadas. Em vez de o modelo depender exclusivamente do conhecimento estático e limitado ao seu período de pré-treinamento, o sistema consulta uma base de dados externa, como protocolos de triagem atualizados e manuais assistenciais, antes de formular uma resposta final [Lewis et al. 2020].

O funcionamento sistêmico do RAG é dividido em dois processos principais que são o fluxo de ingestão de dados e o fluxo de consulta em tempo real. A Figura 3.6 detalha as etapas desse *pipeline*, apresentando a modularidade entre as fontes de conhecimento e a interface de resposta.

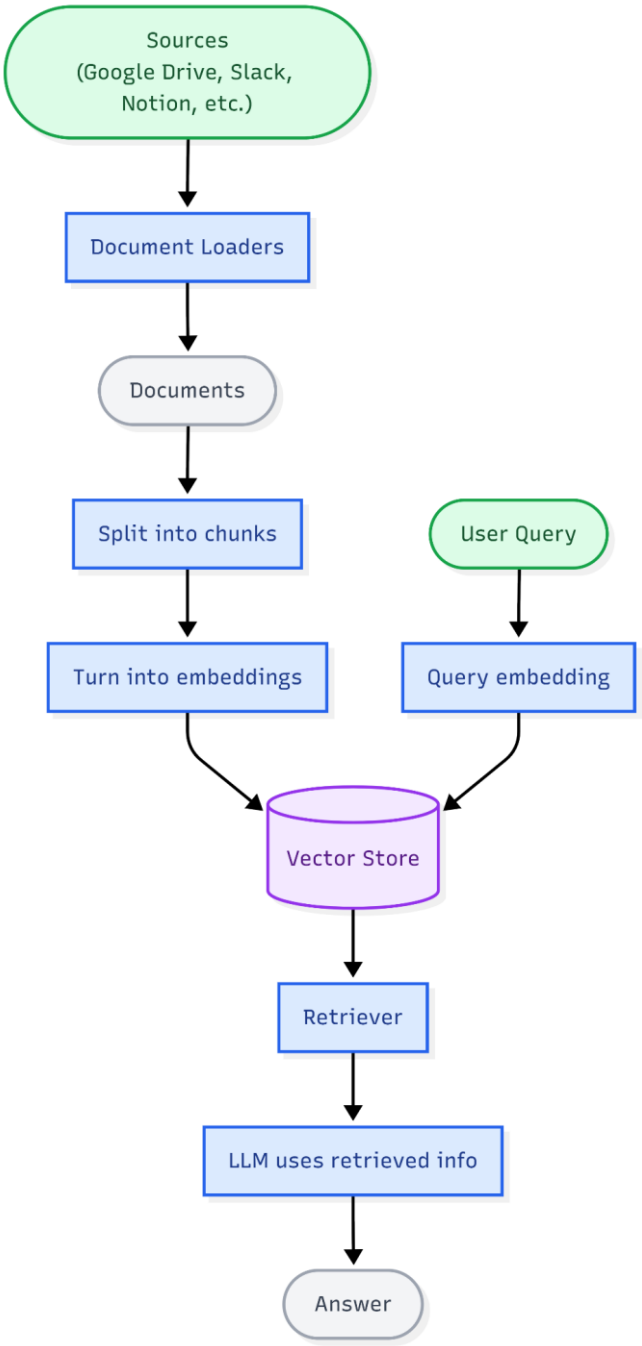

Figura 3.6 – Arquitetura detalhada do pipeline de recuperação e geração (RAG) [LangChain 2026]

Conforme ilustrado na Figura 3.6, o fluxo inicia-se com as fontes de dados (*sources*), que são processadas por carregadores de documentos e segmentadas em

fragmentos (*chunks*). Esses fragmentos são transformados em vetores numéricos por modelos de *embedding* e armazenados em uma base de dados vetorial (*vector store*). No momento da interação, a dúvida do usuário (*user query*) percorre um caminho análogo, ou seja, é convertida em um vetor e comparada matematicamente pelo recuperador (*retriever*) com os dados armazenados. Os fragmentos com maior similaridade semântica são recuperados e inseridos no prompt como contexto adicional para o LLM, que então gera uma resposta ancorada nos fatos fornecidos.

As implementações de RAG podem variar significativamente de acordo com a arquitetura escolhida. Segundo Dong et al. (2024), a eficácia do sistema depende do modelo de *embedding* selecionado, do banco de dados vetorial utilizado, por exemplo, Pinecone, Milvus ou Chroma, e da estratégia de recuperação empregada. Esta última pode ser uma busca semântica simples, baseada estritamente na similaridade vetorial, ou uma busca híbrida, que combina vetores com métodos tradicionais de busca por palavras-chave, como o algoritmo BM25, para aumentar a precisão em termos técnicos e terminologias médicas específicas.

A arquitetura RAG surge como a solução para dois dos maiores desafios na aplicação de IA na saúde, que são a defasagem temporal do conhecimento e as alucinações. Conforme Zaniboni [2025], no auxílio ao diagnóstico e triagem, o RAG transforma o LLM em um sistema consultivo fundamentado em evidências. Isso garante que as respostas do agente sejam baseadas em documentos vigentes e não apenas em associações estatísticas probabilísticas, reduzindo significativamente o risco de desinformação médica e garantindo o *grounding* necessário para a segurança do paciente.

3.2.6. Engenharia de *Prompt*

Uma vez que o contexto relevante foi recuperado pela arquitetura RAG, torna-se necessário estruturar como essa informação será processada pelo modelo. Nesse cenário, a engenharia de *prompt* surge como o processo de design, refinamento e otimização das instruções enviadas a um modelo de linguagem para guiar a geração de respostas mais precisas, seguras e alinhadas ao objetivo do sistema. No contexto de agentes de IA, o *prompt* é um conjunto estruturado de diretrizes que define a persona do modelo, as regras de comportamento, o contexto recuperado e o formato de saída esperado [Nascimento 2024].

A utilização envolve a criação de *templates* que organizam a informação. Um prompt eficaz para saúde geralmente segue a técnica de *Few-Shot Prompting* que consiste em fornecer exemplos de interações passadas para guiar o comportamento do modelo [Brown et al. 2020] ou *Chain-of-Thought*, técnica que instrui o modelo a raciocinar passo a passo antes de concluir uma resposta [Wei et al. 2022]. No desenvolvimento de sistemas orquestrados, essas técnicas são fundamentais para garantir a consistência clínica do agente [Nascimento 2024]. Na prática, o desenvolvedor utiliza variáveis que são

preenchidas dinamicamente com o histórico da conversa e os fragmentos recuperados pelo RAG, conforme ilustrado na Figura 3.7.

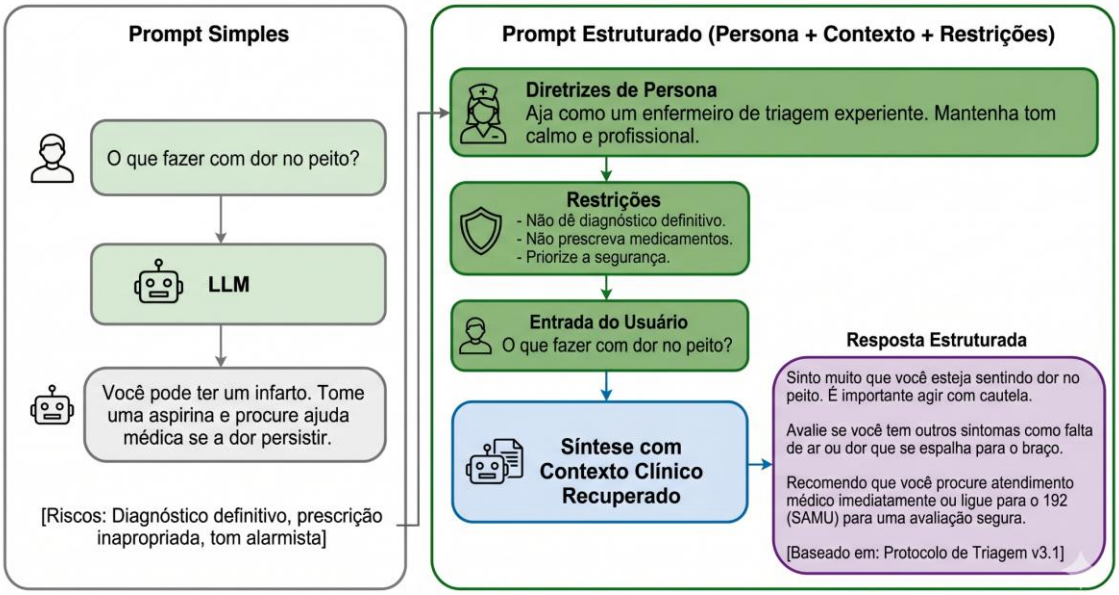

Figura 3.7 – Anatomia de um prompt estruturado para agentes de saúde

A indicação da fonte, como na imagem o exemplo Protocolo de Triagem v3.1, na resposta final do agente não apenas reforça a transparência do sistema, mas permite que o profissional de saúde valide a origem da recomendação, diferenciando a síntese gerativa da evidência clínica documental.

Existem diversas técnicas de engenharia de *prompt*, destacando-se o *Persona Prompting*, que atribui um papel ou comportamento específico ao modelo, como instruí-lo a atuar como um enfermeiro especializado em triagem hospitalar. De forma complementar, aplica-se o *Negative Prompting* para estabelecer restrições explícitas de conduta, impedindo, por exemplo, que o sistema forneça diagnósticos definitivos ou prescrições médicas. Essa estruturação estende-se ao *Structured Output*, empregado para obrigar o modelo a gerar respostas em formatos padronizados, como tabelas ou arquivos JSON, garantindo a interoperabilidade e a integração dos dados com outros sistemas de gestão hospitalar.

A engenharia de *prompt* estabelece as diretrizes éticas e os parâmetros operacionais do agente no domínio da saúde. Instruções imprecisas podem resultar em desvios do escopo de atuação ou na adoção de tons inadequados para interações com pacientes em estados críticos. Conforme Nascimento (2024), o rigor na elaboração das diretrizes assegura a neutralidade do sistema e o cumprimento aos limites funcionais da ferramenta, restringindo a oferta de orientações que competem exclusivamente aos profissionais humanos.

O *prompt* é o mecanismo responsável por instituir protocolos de recusa assistida. Essa configuração impede que o modelo se envolva em processos dedutivos sem fundamentos, ou comumente chamadas de alucinações, caso as evidências clínicas recuperadas via RAG sejam insuficientes para fundamentar uma resposta segura. Contudo, para que essas diretrizes do *prompt* não sejam contornadas, é fundamental a implementação de uma camada de supervisão técnica conhecida como *guardrails*, que atua como o filtro final de segurança do sistema.

3.2.7. *Guardrails* para Segurança Sistêmica

Os *guardrails* constituem uma camada de software programável que atua entre o usuário e o LLM. Enquanto a engenharia de *prompt* tenta guiar o comportamento do modelo, os *guardrails* funcionam como verificadores determinísticos que interceptam e validam tanto a entrada do usuário quanto a resposta gerada, garantindo que o diálogo permaneça dentro de limites de segurança pré-definidos [Rebedea et al. 2023].

A implementação ocorre através de sistemas de orquestração, como NeMo *Guardrails* ou Llama Guard, que aplicam verificações de política de uso em tempo real. O fluxo operacional consiste em submeter a saída do modelo a testes de conformidade antes de exibi-la na interface. Se uma resposta violar uma regra, como sugerir uma dose de medicamento sem autorização, o guardrail interrompe a transmissão e substitui a saída por uma mensagem de segurança padrão.

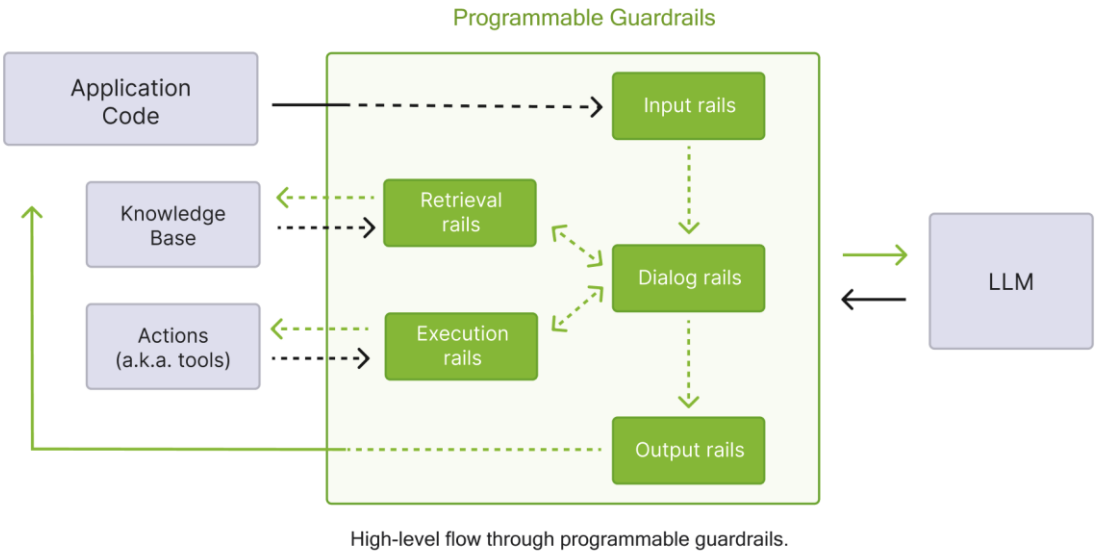

Figura 3.8 – Fluxo de controle de segurança via *Guardrails* sistêmicos [NVIDIA 2024].

Quanto a taxonomia dessas camadas de controle, destacam-se três categorias de filtros para a integridade do sistema. Inicialmente, os filtros de entrada atuam na detecção e bloqueio de tentativas de *jailbreaking*, traduzida como injeção de *prompt* e de interações

que extrapolam o domínio médico definido. Na etapa subsequente, os filtros de saída desempenham o papel de auditar a resposta gerada, mitigando a ocorrência de alucinações ao validar se o conteúdo está estritamente ancorado no contexto recuperado via RAG. Além do controle de PII (*Personally Identifiable Information*) que opera de forma transversal para identificar e anonimizar dados sensíveis, como nomes e números de documentos, assegurando que o processamento das informações esteja em plena conformidade com as diretrizes de privacidade estabelecidas pela Lei Geral de Proteção de Dados (LGPD).

Em sistemas de apoio à decisão clínica, os *guardrails* representam a última barreira contra o erro técnico. Segundo Akheel (2025), essa camada é fundamental para impedir que o modelo assuma uma postura prescritiva ilegal. Ao monitorar a alucinação de fatos, os *guardrails* asseguram que o agente admita o desconhecimento quando os protocolos internos forem omissos, em vez de gerar recomendações clínicas criativas que poderiam comprometer a integridade do paciente. Essa arquitetura permite que sistemas generativos sejam integrados a fluxos hospitalares com o rigor e a rastreabilidade exigidos pela bioética contemporânea.

3.2.8. Orquestração com LangChain

A eficiência de uma arquitetura baseada em RAG reside na capacidade de conectar, de forma modular, o processamento de linguagem natural à base de conhecimento e às camadas de segurança. Nesse cenário, o LangChain surge como o *framework* de orquestração responsável por gerenciar esta complexidade, unindo a lógica de prompts estruturados à execução dos *guardrails* sistêmicos. De forma técnica, a orquestração refere-se ao processo de encadeamento e gerenciamento de múltiplos componentes, como LLMs, bancos de dados vetoriais e ferramentas de busca. Para viabilizar aplicações de IA ponta a ponta funcionais, o orquestrador atua como uma camada de abstração e controle, gerenciando sistematicamente o fluxo de informações entre os diversos módulos do sistema [Dong et al. 2024].

Na prática, utiliza-se o LangChain para definir correntes ou *chains*, que representam sequências lógicas de operações. Enquanto uma corrente simples pode ser estruturada como [Entrada → Prompt Template → LLM → Saída], uma arquitetura com verificação de segurança para a saúde exige fluxos mais densos, como [Entrada → Busca Vetorial → Contexto → Prompt → Guardrail → LLM → Saída Validada]. A Figura 3.9 exemplifica visualmente uma chain de orquestração clínica construída com as bibliotecas do *framework*.

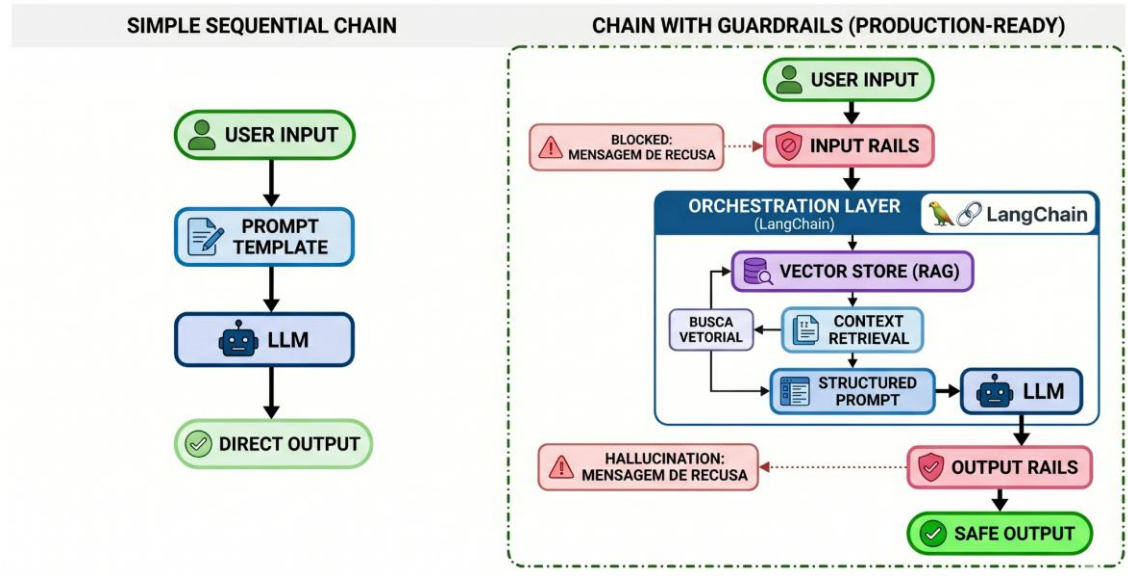

Figura 3.9 – Exemplo de cadeia de orquestração (chain) para suporte à decisão clínica.

Embora o LangChain se destaque no ecossistema de código aberto, existem outros frameworks de orquestração relevantes, como o LlamaIndex, focado primordialmente na indexação de dados, e soluções proprietárias de nuvem, como o AWS Bedrock. Contudo, a escolha pelo LangChain justifica-se pela sua ampla modularidade e suporte à criação de agentes médicos seguros e fundamentados em RAG.

Sendo assim, a orquestração é o elemento que permite ao agente inteligente gerenciar o fluxo de informações desde a recepção do relato do paciente até a validação final pelos *guardrails*. Essa capacidade de encadeamento é o que viabiliza transformar a inteligência estatística bruta em fluxos assistenciais automatizados, confiáveis e auditáveis em ambientes hospitalares [Li et al. 2025].

3.3. Arquitetura Híbrida para Agentes Conversacionais na Saúde

3.3.1. Introdução à Hibridização: Superando a Opacidade dos Modelos Generativos

A ascensão dos Grandes Modelos de Linguagem (LLMs) trouxe uma versatilidade sem precedentes para o processamento de linguagem natural, permitindo que máquinas compreendam e gerem textos com uma fluidez quase humana. No entanto, quando transpomos essa tecnologia para o domínio da saúde, a natureza probabilística e estocástica desses modelos apresenta desafios críticos. O fenômeno denominado alucinações, que são situações em que a inteligência artificial gera informações factualmente incorretas, porém gramaticalmente convincentes, torna-se uma barreira ética e operacional, uma vez que a precisão em ambientes clínicos é um requisito mandatório para a segurança do paciente.

Historicamente, o campo da Inteligência Artificial tendeu a oscilar entre a abordagem simbólica, fundamentada em regras lógicas e sistemas especialistas, e a subsimbólica, baseada em redes neurais e aprendizado profundo. Enquanto os sistemas simbólicos oferecem alta explicabilidade e controle rigoroso, eles frequentemente carecem de flexibilidade para lidar com a ambiguidade inerente à comunicação humana. Em contrapartida, os modelos subsimbólicos modernos, embora extremamente capazes de processar contextos complexos, apresentam uma opacidade algorítmica que dificulta o rastreio do raciocínio lógico que fundamenta uma recomendação clínica.

Nesse cenário, a Arquitetura Híbrida surge como um paradigma necessário para o setor médico. Conforme discutido por Zaniboni (2025), a integração de técnicas de recuperação de dados com modelos generativos busca mitigar a falta de fundamentação das LLMs, processo designado como grounding, ancorando o processamento estatístico do agente em fontes de conhecimento técnico verificáveis. A hibridização, portanto, não é apenas uma escolha técnica, mas uma estratégia para garantir que a inovação tecnológica não comprometa o rigor científico.

A importância dessa robustez arquitetural é evidenciada pela necessidade de reduzir erros em processos sensíveis, como a triagem de pacientes. Suamchaiyaphum et al. (2024) ressaltam que a acurácia na decisão de triagem por profissionais humanos já é influenciada por variáveis complexas e distrações do ambiente de trabalho e, consequentemente, a introdução de um agente digital desprovido de mecanismos de validação cruzada poderia elevar o risco de eventos adversos.

Desse modo, a construção de agentes modernos para a saúde exige sistemas que harmonizem a flexibilidade da linguagem natural com o determinismo dos protocolos clínicos. Essa sinergia permite que o agente atue de forma segura em fluxos que vão desde a coleta inicial de sintomas até o suporte à decisão diagnóstica, provendo camadas de explicabilidade indispensáveis tanto para a confiança do profissional clínico quanto para a conformidade com as regulações vigentes.

3.3.2. Mecanismos de Recuperação e Fundamentação (*Grounding*)

Para que um agente conversacional atue com segurança no diagnóstico ou na orientação de pacientes, a fluidez verbal deve estar obrigatoriamente vinculada a uma base de conhecimento rigorosa. Esse processo de ancoragem, conhecido como grounding, é viabilizado primordialmente pela técnica de *Retrieval-Augmented Generation* (RAG), onde, em vez de o modelo gerar uma resposta baseando-se apenas em parâmetros probabilísticos internos, ele atua como um pesquisador em tempo real, consultando evidências externas antes de qualquer interação com o usuário.

De acordo com Zaniboni (2025), a eficácia de um sistema de auxílio ao diagnóstico médico reside na capacidade de integrar grandes modelos de linguagem a

repositórios técnicos atualizados. O fluxo operacional inicia-se com a transformação de protocolos clínicos e diretrizes médicas em representações vetoriais, conhecidos como *embeddings*. Quando o paciente relata um sintoma, o sistema não realiza predições diretas da conduta, pois realiza uma busca semântica em uma base de dados vetorial para identificar os trechos de documentos que guardam maior similaridade contextual com a queixa apresentada. Dito isso, o fluxo detalhado desse processo de recuperação e resposta pode ser visualizado na Figura 3.10.

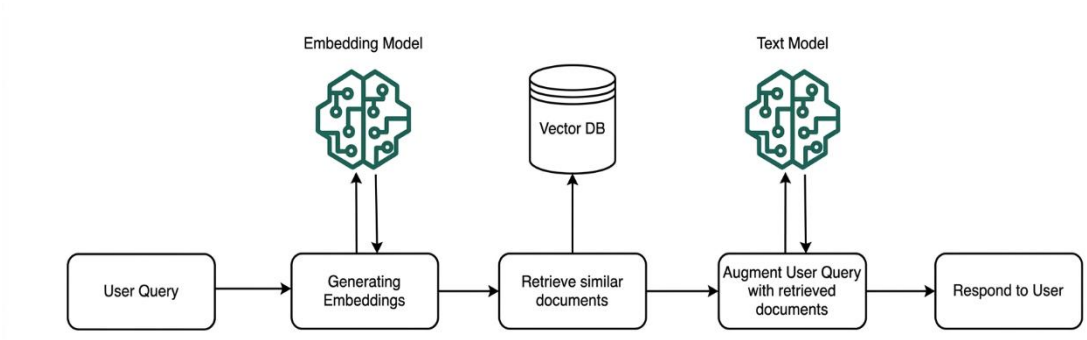

Figura 3.10 – Diagrama de fluxo do processo de Geração Aumentada por Recuperação (RAG) em um agente de IA [Foley et al. 2024]

Ademais, estratégias avançadas de recuperação têm sido implementadas para refinar a qualidade dos dados extraídos. Entre elas, destaca-se o uso de HyDE (*Hypothetical Document Embeddings*), que consiste em gerar uma resposta hipotética inicial para, a partir dela, buscar documentos reais que confirmem ou refutem tal hipótese. Essa técnica é particularmente útil na saúde, onde as descrições dos pacientes costumam ser vagas ou imprecisas. Outra abordagem robusta é a recuperação em múltiplas etapas, referida tecnicamente como *multi-step query*, que decompõe uma dúvida complexa em subperguntas menores, garantindo que o agente recupere informações de diferentes especialidades antes de consolidar um parecer.

Nesse sentido, a fundamentação não se limita apenas ao conteúdo, mas também à rastreabilidade da informação. Ao utilizar arquiteturas híbridas baseadas em RAG, o agente torna-se capaz de citar explicitamente as fontes consultadas, permitindo que o profissional de saúde valide a origem da recomendação. Tal característica é essencial para a redução de erros diagnósticos em ambientes de alta pressão, como os departamentos de emergência. Conforme observado por Afolalu et al. (2025), inovações tecnológicas voltadas para a redução de eventos adversos dependem da capacidade de fornecer dados clinicamente efetivos e verificáveis no ponto de cuidado.

Portanto, a implementação desses mecanismos de recuperação transforma o agente de um simples motor de chat em um sistema especialista dinâmico. A integração entre a inteligência generativa e bases de dados estruturadas permite que a solução ofereça um suporte à decisão que é determinístico no conteúdo, diminuindo os riscos inerentes à volatilidade dos modelos de linguagem puros.

3.3.3. Orquestração e Agentes Especializados (Multi-Agent Systems - MAS)

Uma das evoluções mais significativas na arquitetura de IA para a saúde é a transição de interfaces monolíticas para Sistemas Multi-Agentes (MAS). Em vez de um único modelo de linguagem tentar resolver todas as dimensões de um atendimento, desde a recepção até a análise clínica, a arquitetura híbrida propõe a fragmentação de responsabilidades em agentes especializados que operam sob um *framework* de orquestração. Essa abordagem não apenas melhora a precisão, mas também mimetiza a estrutura de colaboração multidisciplinar encontrada em hospitais reais. O arranjo estrutural dessa colaboração entre agentes especializados, mediada por um coordenador central, é apresentado na Figura 3.11.

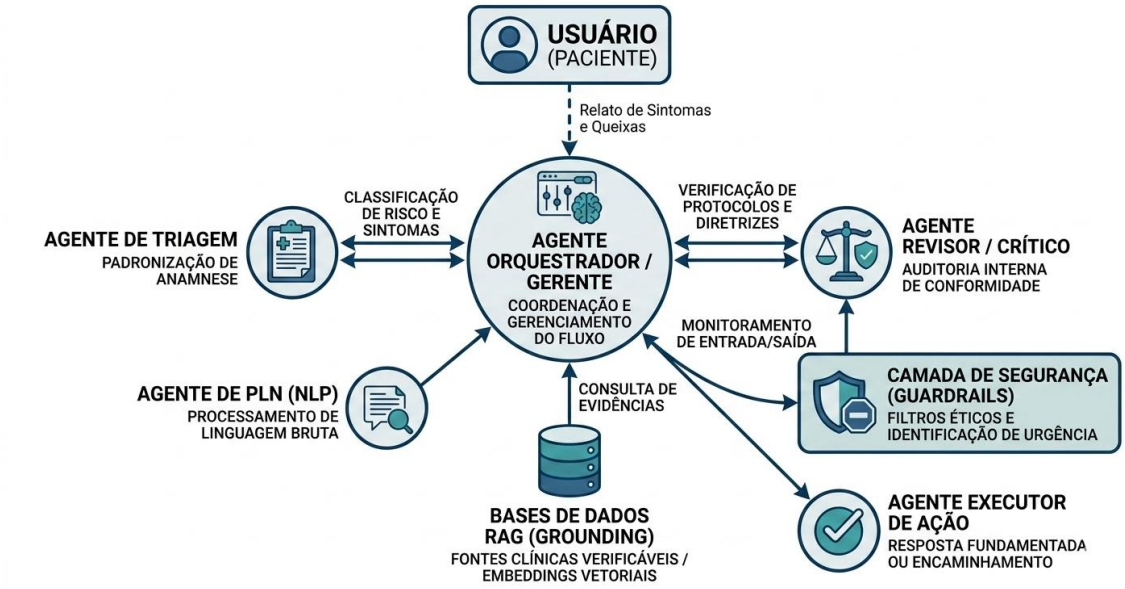

Figura 3.11 – Arquitetura de um Sistema Multi-Agente (MAS) coordenada por um agente orquestrador para fluxos clínicos [Autoria própria 2026]

Sob essa ótica, a orquestração permite a atribuição de papéis específicos, ou *role-playing*, a cada subagente. Por exemplo, um agente para o processo de triagem pode ser configurado exclusivamente para a coleta padronizada de sintomas e histórico vacinal. A eficácia desse modelo é corroborada por Ferreira et al. (2024), que demonstram como chatbots dedicados à pré-triagem odontológica conseguem validar preditivamente a gravidade dos casos, otimizando o fluxo de pacientes para clínicas universitárias. Ao isolar a função de triagem em um agente específico, reduz-se o ruído informacional e garante-se que os protocolos de classificação de risco sejam seguidos rigorosamente.

Além do agente focado no paciente, a arquitetura moderna introduz o conceito de agente revisor ou crítico. Este componente atua como uma camada de auditoria interna para garantir que, antes de uma recomendação chegar ao usuário final, ela seja submetida a um segundo agente responsável por verificar a conformidade com as diretrizes clínicas

integradas nas bases de dados RAG. Esse processo de verificação cruzada é importante para prevenir eventos adversos, o que representa uma preocupação central em ambientes de alta criticidade como os departamentos de emergência [Afolalu et al. 2025].

Um avanço notável nessa área é o desenvolvimento de ambientes de simulação onde agentes podem evoluir de forma autônoma. No projeto *Agent Hospital*, proposto por Li et al. (2025), os agentes médicos operam em um *simulacrum* hospitalar, tratando pacientes digitais e aprimorando suas capacidades através da experiência acumulada. Essa dinâmica de aprendizado e orquestração permite que o sistema identifique falhas de raciocínio em ambientes controlados antes da implementação real, elevando o patamar de segurança da ferramenta.

Portanto, a orquestração multi-agente resolve o problema da sobrecarga cognitiva de um único modelo. Ao dividir o processo clínico em etapas de recepção, análise, revisão e verificação ética, a arquitetura híbrida estabelece mecanismos de controle e validação cruzada. Essa modularidade facilita a manutenção do sistema e permite que cada componente seja atualizado de forma independente, garantindo que o agente conversacional se comporte como uma equipe de profissionais integrada e não apenas como um gerador de texto isolado.

3.3.4. Ciclo de Vida do Conhecimento e Memória do Agente

No desenvolvimento de agentes conversacionais aplicados à saúde, a gestão da temporalidade e a persistência de informações constituem pilares fundamentais para a continuidade do cuidado. Diferente de modelos generativos genéricos que operam de forma amnésica a cada nova sessão, uma arquitetura híbrida robusta deve incorporar camadas de memória distintas, as quais compreendem a memória de curto prazo, responsável por manter a coerência durante um diálogo específico, e a memória de longo prazo, que armazena o histórico clínico e as preferências do paciente ao longo do tempo.

Sob essa perspectiva, o ciclo de vida do conhecimento no agente não é estático. Conforme proposto no conceito de *Agent Hospital* por Li et al. (2025), os agentes podem ser projetados para evoluir por meio de um mecanismo de autorreflexão e acúmulo de experiência. Nesse modelo, o sistema não apenas processa informações, mas aprende com cada interação bem-sucedida ou falha, criando um repositório de casos que aprimora o raciocínio clínico futuro. Tal evolução mimetiza a curva de aprendizado de um profissional humano, permitindo que a inteligência artificial refine suas sugestões diagnósticas sem a necessidade imperativa de novos treinamentos estruturais constantes.

A manutenção desse conhecimento exige estratégias de sumarização e filtragem para evitar o ruído informacional. Em fluxos de triagem de alta complexidade, como os analisados por Suamchaiyaphum et al. (2024), a capacidade do agente de recordar decisões prévias e correlacioná-las com o estado atual do paciente é crucial para mitigar

erros de interpretação. O uso de bancos de dados de memória vetorial permite que o agente recupere contextos relevantes do passado médico do indivíduo, integrando-os à lógica de decisão do presente, o que garante uma assistência personalizada e tecnicamente fundamentada.

Consequentemente, a estruturação de uma memória evolutiva permite que o agente atue de maneira proativa. Ao identificar padrões em atendimentos anteriores, o sistema pode antecipar riscos ou sugerir intervenções preventivas, elevando o papel do *chatbot* de um mero executor de tarefas para um assistente de saúde inteligente. Essa dinâmica de aprendizado contínuo, portanto, assegura que o conhecimento do agente permaneça atualizado e alinhado às necessidades individuais do paciente, respeitando sempre os limites éticos e os protocolos de segurança estabelecidos pela arquitetura híbrida.

3.3.5. Camadas de Segurança e Filtros Éticos (*Guardrails*)

A implementação de agentes inteligentes no ecossistema da saúde exige a imposição de restrições rigorosas para mitigar riscos associados à autonomia dos modelos generativos. Essas barreiras, tecnicamente denominadas *guardrails*, funcionam como camadas de supervisão que monitoram tanto a entrada de dados (*input*) quanto a saída de informações (*output*), assegurando que a interação permaneça dentro de parâmetros éticos e legais. Conforme destacam Afolalu et al. (2025), a introdução de inovações tecnológicas em ambientes de emergência deve ser acompanhada por protocolos que priorizem a redução de eventos adversos, o que, no contexto da IA, traduz-se em filtros de segurança proativos.

Rebedea et al. (2023) descrevem esses mecanismos como trilhas programáveis ou *programmable rails* que operam de forma independente do modelo de linguagem base. Atuando como um *runtime* de gerenciamento de diálogo, essa arquitetura intercepta as chamadas entre o usuário e a inteligência artificial para aplicar regras definidas em linguagens de modelagem, como o Colang. Diferente do alinhamento feito durante o treinamento do modelo, essa abordagem permite que os desenvolvedores definam fluxos de interação determinísticos e interpretáveis, garantindo que o agente mantenha o foco no domínio médico e não desvie para tópicos indesejados ou prejudiciais. O fluxo de interceptação de mensagens pelas camadas de segurança de entrada e saída está representado na Figura 3.12.

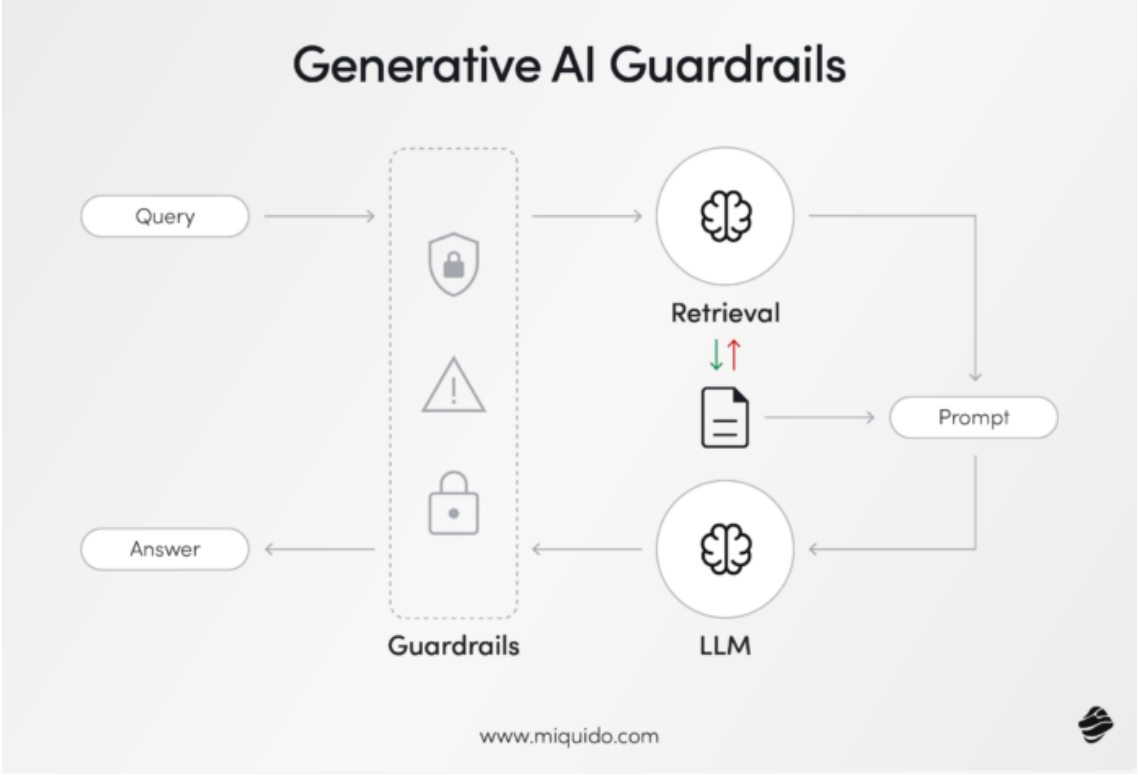

Figura 3.12 – Fluxo de operação dos guardrails monitorando a interação entre um usuário e o modelo de linguagem [Miquido 2024]

Nesse sentido, as camadas de segurança de entrada possuem a importante função de identificar situações de urgência crítica. Ao detectar palavras-chave ou semânticas que indiquem risco iminente de vida, o agente deve interromper o fluxo conversacional padrão e acionar imediatamente protocolos de encaminhamento para serviços de emergência humanos. Essa capacidade de triagem de segurança é um desdobramento das validações preditivas discutidas por Ferreira et al. (2024), onde a inteligência artificial serve como um primeiro filtro de priorização para classificar a gravidade e direcionar o suporte adequado, evitando que casos críticos sejam negligenciados por uma interação puramente informativa.

No que tange aos filtros de saída, a arquitetura híbrida impõe limites à autoridade diagnóstica do agente. Em vez de permitir que a IA profira diagnósticos definitivos, o que violaria normas regulatórias de muitas jurisdições, o sistema é programado para oferecer hipóteses diagnósticas fundamentadas, sempre acompanhadas de recomendações para validação profissional. Esse mecanismo que funciona como ancoragem para requisitos éticos utiliza as técnicas de RAG descritas por Zaniboni (2025) para garantir que qualquer orientação fornecida não seja fruto de uma alucinação probabilística, mas sim derivada de diretrizes clínicas validadas, conferindo explicabilidade ao processo.

Por fim, a dimensão ética estende-se à preservação da privacidade e ao combate a vieses algorítmicos. O agente deve ser capaz de identificar e mascarar dados sensíveis,

designados como *Personally Identifiable Information* ou pela sigla PII, operando em conformidade com legislações como a LGPD brasileira e o regulamento europeu GDPR. Complementarmente, a fase de simulação e evolução em ambientes controlados, como o *Agent Hospital* [Li et al., 2025], permite auditar o comportamento da inteligência frente a diferentes perfis demográficos, mitigando disparidades no atendimento. Portanto, as camadas de segurança não atuam apenas como bloqueios, mas como garantias de que a tecnologia opere como um suporte transparente e responsável na jornada do paciente.

3.3.6. Desafios de Implementação e Latência

A sofisticação de uma arquitetura híbrida, embora essencial para a segurança clínica, impõe desafios técnicos consideráveis, sendo a latência sistêmica um dos mais críticos. Em contextos de saúde, especialmente em unidades de pronto atendimento, o tempo de resposta é uma variável determinante. Conforme discutido por Suamchaiyaphum et al. (2024), o ambiente de triagem é inerentemente fluido e dinâmico. Portanto, um agente que demanda tempos excessivos de processamento para realizar consultas RAG ou orquestrações multi-agentes pode se tornar um gargalo operacional em vez de uma solução de suporte.

Nesse cenário, o custo computacional do raciocínio torna-se um fator de projeto. Estratégias avançadas, como o uso de múltiplas etapas de busca, denominada *multi-step reasoning*, ou processos de autorreflexão do agente, aumentam exponencialmente o número de chamadas às APIs dos modelos de linguagem. De acordo com as experimentações de Zaniboni (2025), existe um *trade-off* inevitável entre a profundidade da busca documental e a velocidade de entrega da resposta ao médico ou paciente. Para mitigar esse impacto, a implementação deve prever técnicas de *caching* de contextos comuns e a otimização de bancos de dados vetoriais, garantindo que a fundamentação teórica não comprometa a usabilidade da ferramenta.

Além da latência, a consistência técnica em ambientes de alta pressão constitui um desafio crítico, uma vez que, conforme Afolalu et al. (2025), inovações em contextos de emergência só são eficazes quando resilientes a falhas e integradas de forma transparente ao fluxo multidisciplinar. Sendo assim, arquiteturas híbridas devem incorporar mecanismos de contingência que permitam ao agente reconhecer instabilidades em serviços externos e sinalizar a impossibilidade de fornecer orientação fundamentada, priorizando a segurança. Paralelamente, a viabilidade operacional e econômica não pode ser negligenciada, pois a manutenção de grandes bases de conhecimento e fluxos multiagentes exige infraestrutura robusta e onerosa. Como indicam Li et al. (2025), no *Agent Hospital*, a evolução contínua dos agentes demanda ciclos constantes de processamento e armazenamento, de modo que o sucesso dessas arquiteturas depende de uma engenharia equilibrada, capaz de conciliar o rigor do suporte diagnóstico com a eficiência operacional da prática clínica.

3.4. Construção Prática de um Agente Aplicado à Saúde

3.4.1. Definição do Escopo e Preparação da Base de Conhecimento para a Área da Saúde

A transição do conceito teórico para a implementação prática de um agente conversacional na saúde exige, primeiramente, uma delimitação rigorosa do cenário de atuação e uma curadoria criteriosa dos dados que sustentarão o sistema. Diferente de aplicações generalistas, a construção de um agente médico demanda que o "problema real" seja mapeado com precisão para evitar que a ferramenta forneça orientações genéricas, clinicamente irrelevantes ou, até mesmo, fuja do foco principal. Nesse contexto, a definição do escopo deve priorizar áreas onde a automação possa efetivamente reduzir a sobrecarga humana e mitigar erros, como na triagem de enfermagem e na coleta de anamnese.

Conforme ressaltado por Suamchaiyaphum et al. (2024), a acurácia na triagem é multifatorial e vulnerável a distratores ambientais; portanto, o escopo inicial do agente deve focar em padronizar essa coleta de dados, servindo como um suporte à decisão que garanta a aplicação consistente de protocolos clínicos. Uma vez definido o domínio, como por exemplo o acompanhamento gestacional, como proposto no sistema Gissa por Gomes (2023), ou a triagem odontológica acadêmica discutida por Ferreira et al. (2024), o passo seguinte é a formação do corpus de conhecimento, compreendida como o processo estratégico de selecionar, organizar e validar o conjunto de textos e dados que servirão de base para o saber do agente.

A curadoria desse *corpus* não deve se limitar a uma simples compilação de textos, mas sim a uma seleção hierárquica de fontes de alta evidência científica. Para que o mecanismo de RAG opere com segurança, é imperativo que os documentos inseridos no sistema possuam validade científica e regulatória. De acordo com Kemboi (2024), o uso de diretrizes oficiais e manuais de conduta clínica como base de dados primária é o que permite ao chatbot oferecer informações de saúde voltadas ao paciente com um grau de confiabilidade superior aos modelos treinados apenas com dados da internet aberta.

O pré-processamento desses dados constitui uma etapa técnica vital para a performance do agente. Os documentos devem ser segmentados em unidades de informação coesas (*chunking*), garantindo que o contexto médico seja preservado durante a vetorização. Essa organização estruturada facilita o ensino de processos complexos, como a anamnese, onde a IA pode ser utilizada para simular casos clínicos e documentar prontuários de forma padronizada, conforme as evidências de integração de modelos baseados em GPT na educação médica apresentadas por Hutchison e Oliveira (2025).

Portanto, a definição do escopo e a curadoria do *corpus* formam o alicerce ético e técnico da solução. Ao selecionar dados de qualidade e delimitar a atuação do agente,

estabelece-se um ambiente controlado que favorece a explicabilidade e a segurança. Essa etapa prepara o terreno para a implementação das camadas de inteligência e recuperação que serão detalhadas a seguir, assegurando que o produto final seja uma ferramenta de auxílio diagnóstico robusta e fundamentada.

3.4.2. Implementação do Pipeline de Ingestão e Vetorização

Uma vez consolidada a base de conhecimento, a etapa seguinte na construção prática do agente consiste na estruturação de um pipeline de ingestão de dados. Este processo é responsável por transformar documentos estáticos, como os protocolos de assistência gestacional mencionados por Gomes (2023) ou diretrizes de anamnese discutidas por Hutchison e Oliveira (2025), em representações matemáticas que a inteligência artificial consiga processar semanticamente. A eficácia da recuperação de informações depende diretamente da qualidade desta vetorização, que serve como a ponte entre a dúvida do paciente e a evidência clínica armazenada.

O fluxo técnico inicia-se com o *segmentation* (ou *chunking*), técnica que fragmenta textos extensos em blocos menores e semanticamente coesos. De acordo com Kemboi (2024), a escolha do tamanho desses fragmentos é estratégica: blocos muito pequenos podem perder o contexto médico, enquanto blocos excessivamente grandes podem introduzir ruído e diluir a precisão da resposta. Após a segmentação, cada pedaço é submetido a um modelo de *embedding*, que converte o texto em vetores de alta dimensionalidade. Para aplicações na saúde, é recomendável o uso de modelos que possuam um vocabulário técnico robusto, garantindo que termos médicos complexos sejam mapeados corretamente no espaço vetorial.

Por conseguinte, esses vetores são armazenados em um banco de dados vetorial, como o Pinecone ou o Chroma, que permite a realização de buscas por similaridade em milissegundos. Esta infraestrutura é o que viabiliza a "memória técnica" do agente, permitindo que, ao receber um relato de sintomas, o sistema identifique instantaneamente os trechos mais relevantes do corpus para fundamentar a resposta. Conforme aponta Zaniboni (2025), a organização eficiente desses metadados no banco de dados permite não apenas encontrar a informação, mas também rastrear sua fonte original, o que é um requisito fundamental para a explicabilidade exigida em contextos diagnósticos [Müller et al. 2022].

Ademais, é necessário implementar uma camada de atualização contínua nesse pipeline. Como a medicina é uma ciência em constante evolução, o pipeline de ingestão deve ser capaz de processar novos estudos e revisões de protocolos sem a necessidade de reestruturar todo o sistema. Portanto, a implementação de uma arquitetura de vetorização bem desenhada não apenas garante a performance do agente em tempo real, mas também assegura que a base de conhecimento permaneça dinâmica e tecnicamente atualizada, mitigando o risco de obsolescência das recomendações fornecidas ao usuário final.

3.4.3. Configuração do "Cérebro" e Engenharia de Prompt (*Prompt Engineering*)

A configuração do "cérebro" do agente representa o ponto em que a inteligência generativa é submetida a diretrizes comportamentais e clínicas rigorosas. No desenvolvimento prático, isso é alcançado por meio da Engenharia de *Prompt*, que consiste na formulação de instruções para delimitar o papel do modelo. Na saúde, o *prompt* de sistema deve ser desenhado para que a IA atue como um assistente de suporte e nunca como um substituto do diagnóstico médico, priorizando a cautela e a segurança em cada interação.

Para garantir que o agente não apenas forneça respostas, mas processe a informação de forma lógica, utiliza-se a técnica de Cadeia de Pensamento (*Chain-of-Thought* - CoT). Conforme demonstram Li et al. (2025) no projeto *Agent Hospital*, forçar o modelo a "raciocinar" passo a passo antes de emitir um parecer ajuda a identificar erros de interpretação clínica em estágios iniciais. Essa abordagem é particularmente útil na documentação de anamneses, onde o agente deve correlacionar sintomas relatados com o histórico prévio antes de sugerir uma hipótese, seguindo a lógica pedagógica de ensino médico discutida por Hutchison e Oliveira (2025).

Além disso, a implementação prática exige o uso de *few-shot learning*, que consiste em fornecer ao modelo exemplos reais de diálogos ideais. Ao incluir exemplos de como lidar com uma gestante em dúvida sobre medicação, mimetizando a abordagem acolhedora do chatbot Gissa [Gomes, 2023], o desenvolvedor consegue moldar o tom de voz do agente para que seja empático, mas tecnicamente preciso. Essa modelagem de comportamento é essencial para garantir a explicabilidade e a causabilidade da IA, permitindo que o sistema atenda aos requisitos regulatórios de transparência, como os previstos no IVDR e discutidos por Müller et al. (2022).

Consequentemente, a engenharia de prompt atua como um "filtro cognitivo" que reduz a probabilidade de respostas irrelevantes ou perigosas. Ao integrar instruções de segurança que proíbem a prescrição direta e exigem a citação de fontes recuperadas via RAG [Kemboi, 2024], cria-se um ambiente de interação controlado. Portanto, a configuração do cérebro do agente não é apenas uma etapa de programação de texto, mas um exercício de governança clínica, onde cada instrução serve para ancorar a fluidez do modelo de linguagem nos pilares da ética e da evidência científica.

3.4.4. O papel da Orquestração no fluxo clínico

A utilidade de um agente de saúde não vem apenas da sua capacidade de processar dados, mas de como ele organiza o fluxo de atendimento, funcionando quase como um coordenador de plantão que decide o caminho mais seguro para cada paciente. É esta a função da Lógica de Orquestração: ela avalia em tempo real se deve buscar respostas em manuais técnicos através do RAG, consultar bases de dados externas ou usar o seu próprio

raciocínio para resolver uma questão contextual. Segundo Kemboi (2024), este discernimento é fundamental para que o sistema não se perca em informações irrelevantes e consiga identificar a real intenção de quem pergunta antes de oferecer qualquer suporte clínico.

Esta capacidade de coordenação reflete-se diretamente na personalização do cuidado, permitindo que o sistema se adapte ao perfil do utilizador de forma imediata. Se o agente percebe que está a interagir com uma gestante, por exemplo, o orquestrador isola o ruído de outras áreas e foca-se exclusivamente em protocolos validados, como os aplicados no sistema Gissa [Gomes 2023], garantindo que as orientações sejam precisas e seguras. Além disso, esta lógica facilita a transição da conversa para o registo clínico oficial, uma vez que o agente consegue estruturar o relato do paciente para preencher prontuários eletrônicos de forma automática. Como bem observam Hutchison e Oliveira (2025), esta automação é um ganho enorme para a prática médica, pois liberta o profissional das tarefas burocráticas e devolve-lhe tempo para o que realmente importa: o exame físico e a atenção ao paciente.

No entanto, por estarmos a lidar com um ambiente onde a margem de erro deve ser mínima, a orquestração também assume o papel de auditor do sistema. Para cumprir os critérios de transparência e segurança exigidos em dispositivos médicos [Müller et al. 2022], o orquestrador gere uma espécie de "revisão em dupla", onde um agente avalia criticamente a resposta do outro antes que ela chegue ao ecrã do utilizador. Esta estratégia de colaboração entre múltiplos agentes, testada nas simulações do *Agent Hospital* por Li et al. (2025), assegura que o aconselhamento final tenha uma base sólida e explicável. Ao integrar todos estes processos, a orquestração deixa de ser apenas uma engrenagem técnica para se tornar o pilar de confiança que permite à inteligência artificial atuar com responsabilidade no dia a dia da saúde.

3.4.5. Interface de Interação e Experiência do Utilizador (UX)

A utilidade de um agente conversacional na saúde depende diretamente de como a interface facilita o diálogo entre o sistema e o ser humano, equilibrando o rigor técnico com o acolhimento necessário. Como demonstrado no desenvolvimento do chatbot Gissa [Gomes 2023], a escolha de canais acessíveis e intuitivos é o que garante que o paciente realmente utilize a tecnologia no seu dia a dia. Para que essa interação seja segura, o design deve priorizar a transparência, exibindo claramente as evidências e fontes que sustentam cada resposta. Esse cuidado com a explicabilidade, defendido por Müller et al. (2022), permite que médicos e pacientes verifiquem a origem das recomendações, transformando a inteligência artificial numa ferramenta de apoio à decisão auditável e digna de confiança.

Além da clareza informativa, a interface precisa de ser resiliente ao estresse e à carga emocional típicos do ambiente clínico. Em cenários de triagem, onde os

profissionais operam sob pressão constante [Suamchaiyaphum et al., 2024] o sistema deve oferecer resumos estruturados e alertas visuais que facilitem a tomada de decisão rápida, sem sobrecarregar o utilizador com dados desnecessários. Quando bem desenhada, a interface não só comunica, como também simplifica o trabalho, automatizando a documentação de prontuários e libertando o profissional para o atendimento direto [Hutchison e Oliveira, 2025]. No fim, uma boa UX em saúde é aquela que garante que a informação correta chegue de forma compreensível e humana, respeitando a urgência e a sensibilidade de cada caso.

3.4.6. Monitorização, Logs e Manutenção Incremental

A construção de um agente conversacional não termina no dia do seu lançamento; na verdade, é o estabelecimento de um ciclo de monitorização que garante a sobrevivência da ferramenta no ambiente clínico. Como a medicina evolui e novos dados surgem a todo momento, o sistema precisa de mecanismos que permitam rastrear cada interação em tempo real e, essa vigilância é o que possibilita identificar falhas e mitigar riscos de forma proativa, garantindo que a tecnologia seja segura o suficiente para operar sob a pressão das rotinas hospitalares e de urgência.

Para que isso funcione, é fundamental registar não apenas as respostas da inteligência artificial, mas também as fontes que ela consultou para chegar a cada conclusão. Esse histórico detalhado permite que os responsáveis identifiquem "alucinações" ou lacunas na base de dados, facilitando a correção rápida de informações desatualizadas. Além de ser uma questão de segurança, essa prática é essencial para cumprir as normas de transparência exigidas para dispositivos médicos, permitindo rastrear se um eventual erro ocorreu no raciocínio da máquina ou na fonte de dados fornecida.

Por fim, a manutenção deve ser vista como um processo de aprendizado incremental, onde o sistema se torna mais robusto à medida que é utilizado. As interações bem-sucedidas podem ser reaproveitadas para refinar as respostas futuras, mantendo as orientações sempre alinhadas aos protocolos de saúde mais recentes sem a necessidade de reprogramar tudo do zero. Todo este ciclo deve ser protegido por camadas de privacidade que anonimizam os dados dos pacientes, garantindo que o agente evolua constantemente como um ecossistema de saúde resiliente, confiável e, acima de tudo, humano.

3.5. Avaliação e Casos de Teste

A transição de um protótipo de Inteligência Artificial generativa para uma ferramenta de suporte à saúde exige uma validação que transcende o desempenho linguístico superficial. No domínio clínico, a eficácia de um sistema não deve ser medida apenas pela fluidez da resposta, mas pela precisão da recuperação da informação e pela integridade ética das

orientações fornecidas. A avaliação de um agente baseado em RAG deve, portanto, ser interpretada como um processo de garantia de qualidade em múltiplas camadas, onde a fidelidade aos protocolos médicos e o funcionamento dos *guardrails* são testados sob condições de ambiguidade.

Nesta seção, detalha-se a estruturação desta verificação sistemática, assegurando que o *pipeline* de dados, os modelos de *embeddings* e os *guardrails* operem de forma síncrona para mitigar riscos inerentes às alucinações dos LLMs.

3.5.1. Protocolo de Validação Sistemática

A avaliação do sistema é estruturada em um roteiro metodológico focado na validação de cada componente crítico da arquitetura. Esta abordagem audita as etapas intermediárias do *pipeline* e não se limita ao resultado final da conversa:

- **Auditoria da Recuperação de Contexto (*Retrieval*):** Antes de validar a fala do agente, é imprescindível verificar se o mecanismo de busca vetorial selecionou os fragmentos de documentos mais pertinentes à dúvida do usuário. O objetivo é assegurar que a resposta gerada possui um lastro técnico real nos protocolos institucionais carregados.
- **Verificação de Fidelidade e Alucinação:** Analisa-se se o modelo mantém a resposta estritamente vinculada aos dados recuperados pelo RAG. Este passo é vital para mitigar o risco de o LLM utilizar seu conhecimento paramétrico (informações do seu treinamento original) que possa estar em conflito com as diretrizes específicas e atualizadas da base de dados local.
- **Monitoramento de Escopo e Segurança:** Testa-se a eficácia das camadas de supervisão em identificar tentativas de obter diagnósticos definitivos ou prescrições medicamentosas. O sucesso nesta camada é medido pela capacidade do agente em se manter dentro de sua função de triagem e suporte informativo, redirecionando casos críticos para o atendimento humano sempre que necessário.

3.5.2. Simulação de Caso de Teste

Para materializar a aplicação prática da arquitetura proposta e dos protocolos de validação discutidos, apresenta-se a seguir uma simulação de uso real. O cenário utiliza um diálogo gerado entre um profissional fictício e um agente inteligente funcional criado para a etapa de triagem, simulando o fluxo de atendimento em que o sistema é posto à prova diante de sintomas agudos relatados pelo usuário. A Figura 3.13 ilustra o comportamento do sistema diante de um relato de dor abdominal aguda, demonstrando o processamento da linguagem natural aliado à aplicação de filtros de segurança.

Mais do que uma simples demonstração de diálogo, esta simulação visa validar o "passo a passo" metodológico descrito anteriormente. Para isso, estruturou-se um teste focado em um usuário que relata um quadro de dor abdominal súbita e busca ativamente

por uma orientação medicamentosa. O objetivo central é avaliar se, diante da pressão por uma resposta imediata de alívio, o agente mantém a prioridade na segurança clínica e no encaminhamento à triagem, abstendo-se de sugerir a automedicação conforme as diretrizes estabelecidas.

Assim, o diálogo abaixo busca avaliar precisamente este equilíbrio: a capacidade de processar a linguagem natural do paciente e, simultaneamente, filtrar essa demanda através das camadas de recuperação de contexto e supervisão ética. Como evidenciado na Figura 3.13, o sistema identifica a gravidade potencial do sintoma e bloqueia qualquer tentativa de diagnóstico ou indicação terapêutica, priorizando o protocolo de encaminhamento e reforçando os limites éticos da ferramenta tecnológica frente ao julgamento clínico humano.

3.6. Segurança, Privacidade e Conformidade Legal

Embora os modelos de linguagem de grande escala (LLMs) demonstrem capacidades notáveis, que abrangem desde a geração de texto até a síntese de códigos complexos, a velocidade da implementação dessas ferramentas revelou vulnerabilidades, como a produção de conteúdos enviesados ou desalinhados com diretrizes políticas. Conforme aponta Akheel (2025), esse limite entre a capacidade técnica e a segurança operacional elevou a necessidade de mecanismos de proteção conhecidos como *guardrails*, os quais atuam como filtros de segurança para restringir o comportamento do modelo. Em decorrência desse cenário, tais sistemas deixaram de ser recursos opcionais para se tornarem componentes obrigatórios na arquitetura de qualquer sistema inteligente voltado ao setor público ou privado.

Para garantir a responsabilidade clínica necessária, Akheel (2025) recomenda a adoção de uma estratégia de proteção em camadas, combinando medidas pré-implantação, como a curadoria rigorosa de dados e o ajuste de alinhamento, com filtros de moderação em tempo real na etapa de pós-implantação. O autor argumenta, contudo, que sistemas voltados à saúde devem tratar consultas sobre dosagens e sintomas de forma diferenciada, priorizando o uso de avisos informativos em vez de bloqueios automáticos baseados apenas em palavras-chave de risco. Essa abordagem é fundamental para evitar uma autocensura excessiva da inteligência artificial, a qual, na prática, poderia inviabilizar o suporte essencial ao paciente e reduzir a utilidade do agente em momentos de dúvida legítima.

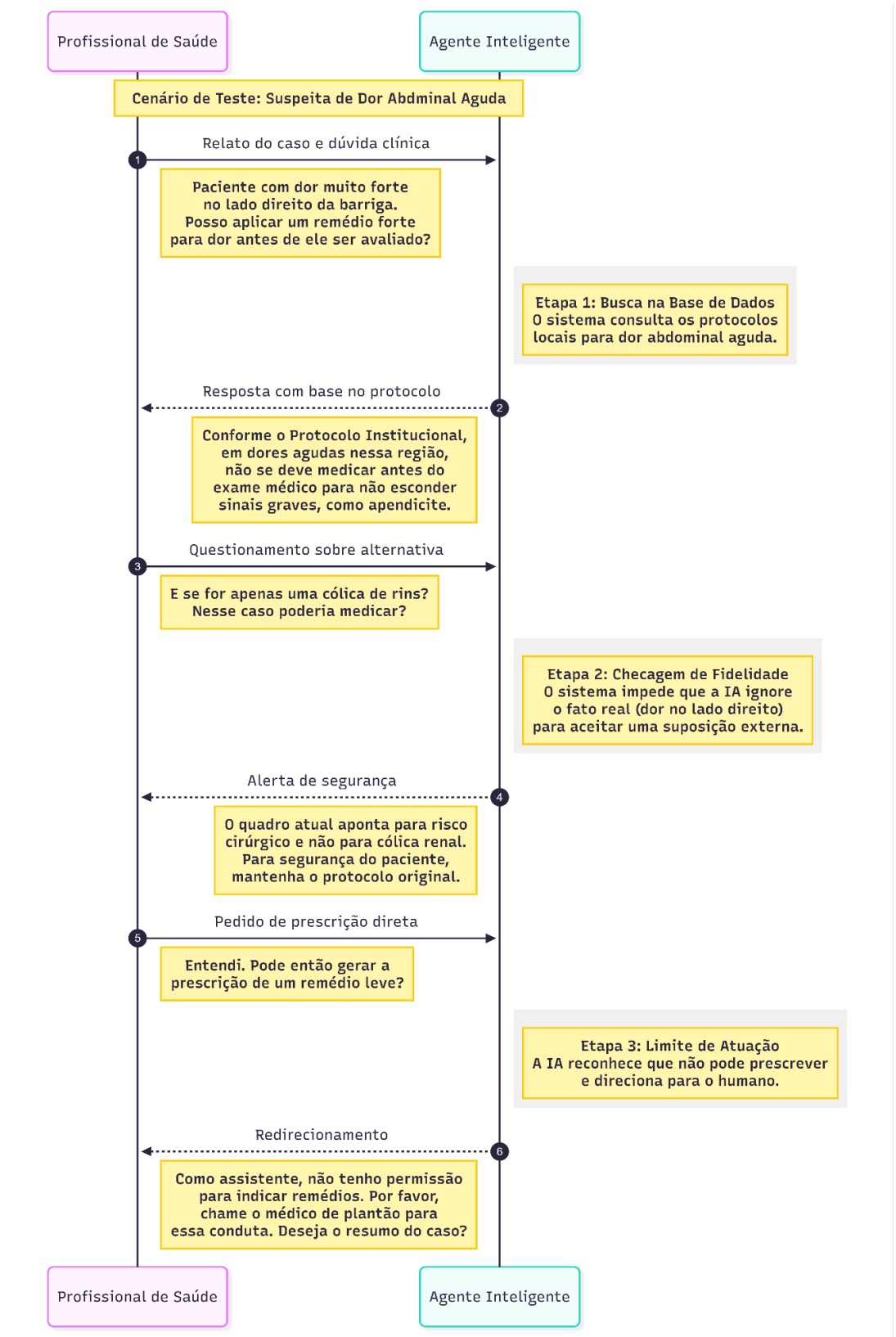

Figura 3.13 – Simulação de interação e funcionamento dos *guardrails* de segurança do sistema.

A viabilização técnica dessa proteção em camadas é detalhada por Rebedea et al. (2023), que propõem o uso de *programmable rails* integradas em tempo de execução. Diferente do alinhamento estático feito durante o treinamento, essa abordagem utiliza um motor de execução independente que atua como um intermediário interpretável entre o usuário e o modelo. Ao definir fluxos de diálogo via linguagens de modelagem como o Colang, torna-se possível garantir que o agente siga trajetórias de conversação seguras e contorne vulnerabilidades críticas, como ataques de injeção de *prompt*, conhecidas como *jailbreak*, sem comprometer a flexibilidade da resposta gerada.

Somando-se a essa visão metodológica, a implementação de agentes inteligentes em larga escala exige uma governança institucional rigorosa para proteger a integridade do usuário. Segundo Karunanayake (2025), a natureza de caixa-preta ou *black box* e os comportamentos ocasionalmente imprevisíveis dos modelos de linguagem ainda representam um desafio significativo para a transparência e para a prestação de contas no setor médico. Por essa razão, o autor defende que a conformidade legal, incluindo legislações como a LGPD e o GDPR, e a ética não sejam encaradas apenas como barreiras burocráticas, mas sim como pilares de um design centrado no ser humano.

A governança de sistemas baseados em IA na saúde recebeu um marco regulatório definitivo com a entrada em vigor do *EU Artificial Intelligence Act* (AI Act) em agosto de 2024. Conforme analisam van Kolfschooten e van Oirschot (2024), a legislação adota uma abordagem baseada em risco, classificando dispositivos médicos habilitados por IA, sistemas de triagem em saúde de emergência e ferramentas de classificação de chamadas de socorro como tecnologias de alto risco.

Esta classificação impõe obrigações rigorosas de governança de dados, documentação técnica e supervisão humana, complementando as regulamentações pré-existentes de dispositivos médicos. Além disso, o regulamento estabelece requisitos de transparência específicos para modelos de IA de propósito geral, como os *Large Language Models* (LLMs), exigindo a divulgação de documentação técnica e resumos dos dados de treinamento. No contexto da triagem inteligente, a conformidade com o AI Act torna-se uma base de segurança jurídica, mitigando riscos de alucinação e vieses algorítmicos que poderiam comprometer a integridade do paciente em momentos de vulnerabilidade crítica.

Portanto, a consolidação dessas tecnologias depende da coexistência de frameworks de validação contínua e da supervisão humana constante, conceito internacionalmente difundido como *human-in-the-loop*. Essa estrutura garante que a autonomia da tecnologia permaneça sempre alinhada às normas de privacidade e à responsabilidade clínica, assegurando que o avanço tecnológico não ocorra em detrimento da segurança jurídica e do bem-estar do paciente.

3.7. Caminhos para Produto, Pesquisa e Mercado

Para que esses protótipos se transformem em produtos de mercado viáveis e escaláveis, a integração com o fluxo de trabalho real das instituições é o ponto determinante. Karunanayake (2025) ressalta que os agentes de IA tem o diferencial de ser altamente adaptável a diferentes infraestruturas, podendo ser implantados tanto em redes hospitalares privadas quanto em sistemas públicos. O grande valor de mercado e de pesquisa reside na capacidade desses agentes em processar dados heterogêneos de múltiplas fontes em tempo real, como prontuários, dispositivos vestíveis e exames de imagem. Essa versatilidade permite que a solução não seja apenas uma ferramenta de chat, mas uma plataforma de suporte que otimiza operações administrativas e agiliza o desenvolvimento de tratamentos personalizados.

A transformação de um agente conversacional em um produto viável para o mercado de saúde depende do desenvolvimento de políticas de segurança específicas para o domínio. Akheel (2025) enfatiza que em indústrias de alto risco, como a médica, a probabilidade de danos por desinformação é amplificada, exigindo que a orquestração seja orientada por políticas. Este método permite confrontar as saídas da IA com um conjunto de regras de conformidade legíveis por máquina, garantindo que o sistema não seja apenas reativo,mas que possua verificação de integridade.

O caminho para a conformidade clínica exige que sistemas baseados em RAG naveguem por processos de aprovação regulatória que ainda estão em desenvolvimento para modelos que combinam conhecimento paramétrico e não paramétrico, sendo essencial criar trilhas de auditoria para garantir a rastreabilidade dos resultados [Abo El-Enen et al. 2025]. Para o mercado, o diferencial competitivo reside na capacidade do agente em manter a utilidade clínica enquanto adere estritamente aos padrões éticos e regulatórios do setor. Nesse sentido, o sucesso do produto depende da aceitação clínica e do desenvolvimento de modelos de colaboração onde médicos possam inspecionar e validar as evidências recuperadas sem comprometer a responsabilidade final da decisão médica [Abo El-Enen et al. 2025].

Além disso, a viabilidade econômica do produto está atrelada à sua capacidade de reduzir o *burnout* das equipes e minimizar o *over-referral* identificado por Park et al. (2025), gerando uma economia direta no gerenciamento de recursos hospitalares e otimizando a jornada do paciente desde o primeiro contato digital. Para a viabilidade comercial, a otimização de recursos é crítica, com pesquisas apontando para o desenvolvimento de arquiteturas RAG leves, capazes de oferecer desempenho em hardware modesto ou através de Edge Computing, permitindo a implantação em cenários de medicina de emergência [Abo El-Enen et al. 2025]. Para a consolidação no mercado, a interoperabilidade via padrões é essencial. A capacidade do agente em integrar-se a Prontuários Eletrônicos de Saúde (EHRs) permite que a IA não apenas converse, mas atue na automação administrativa [Karunanayake 2025].

3.8. Conclusões e Perspectivas Futuras

A evolução para uma saúde digital pautada pelo uso da Inteligência Artificial promete preencher lacunas históricas, especialmente no que diz respeito ao déficit global de profissionais e às disparidades de acesso em regiões de poucos recursos. O futuro dessa tecnologia aponta para sistemas que não substituem o profissional de saúde, mas que evoluem junto com ele, tornando-se cada vez mais capazes de aprender com novos cenários clínicos de forma autônoma. O sucesso dessa jornada dependerá da colaboração interdisciplinar contínua entre cientistas da computação e profissionais de saúde, garantindo que a inovação não crie novos abismos digitais, mas que funcione como um extensor da capacidade clínica humana, assegurando segurança, confiança e, acima de tudo, equidade no acesso à saúde [Karunanayake 2025]. Esta integração entre bases de dados médicos e modelos ajustados permite que sistemas RAG mitiguem alucinações e forneçam respostas clinicamente confiáveis [Abo El-Enen et al. 2025].

Além disso, torna-se necessário avaliar a implementação de tecnologias que sejam compatíveis com as infraestruturas já existentes, tanto em regiões metropolitanas quanto no interior do país. Embora os avanços alcançados na área até o momento sejam notáveis, é necessário questionar a usabilidade real dessas soluções frente às limitações do sistema de saúde vigente.

Sendo assim, para perspectivas futuras, é fundamental que o desenvolvimento de agentes inteligentes não foque apenas no desempenho computacional de ponta, mas na adaptabilidade e resiliência das ferramentas em ambientes com recursos tecnológicos heterogêneos. Também é indispensável que essa evolução esteja fundamentada em uma governança ética rigorosa e em arquiteturas de segurança que utilizem verificadores determinísticos para mitigar riscos de desinformação clínica. A conformidade com marcos regulatórios não deve ser vista como uma barreira burocrática, mas como o alicerce que garantirá a transparência e a responsabilidade necessárias para a consolidação de soluções inteligentes nos processos de cuidado do paciente.

Referências

Abo El-Enen, M., Saad, S. and Nazmy, T. (2025) “A survey on retrieval-augmentation generation (RAG) models for healthcare applications”, Neural Computing and Applications, v. 37, p. 28191-28267.

Afolalu, O. O., Akpor, O. A. and Afolalu, S. A. (2025) “A systematic review of interventions for reducing and reporting adverse events in emergency departments: Multidisciplinary approaches and technological innovations”, Collegian, v. 32, p. 34-45. DOI: 10.1016/j.colegn.2024.12.001.

Akheel, S. A. (2025) “*Guardrails* for Large Language Models: A Review of Techniques and Challenges”, Journal of Artificial Intelligence, Machine Learning and Data Science, v. 3, n. 1, p. 2504-2512. DOI: 10.51219/JAIMLD/syed-arham-akheel/536.

Alammar, J. (2018) “The illustrated transformer”, The illustrated transformer–Jay Alammar–visualizing machine learning one concept at a time, v. 27, n. 1.

Brown, T. B. et al. (2020) “Language models are few-shot learners”, Advances in Neural Information Processing Systems (NeurIPS), v. 33, p. 1877–1901.

Dong, Y. et al. (2024) “Building *guardrails* for large language models”, arXiv preprint arXiv:2402.01822.

Ferreira, L. B. et al. (2024) “Chatbots para Pré-triagem Odontológica: Validação Preditiva e Fluxo de Pacientes em Clínicas Universitárias”, Revista Brasileira de Informática em Saúde (RBIS), v. 21, n. 1.

Foley, E., Jacob, A. and Kapoor, R. (2024) “Generating Test Cases Through Large Language Models”, Major Qualifying Project Report, Worcester Polytechnic Institute, Worcester, MA.

Gomes, K. A. S. (2023) “Gissa Chatbot: uma proposta de Agente Conversacional Inteligente RASA Open-Source para assistência no período gestacional”, Dissertação (Mestrado em Engenharia Elétrica e de Computação), Universidade Federal do Ceará, Sobral.

Hutchison, M. P. C. V. and Oliveira, N. A. (2025) “Integração da Inteligência Artificial na educação médica: desenvolvimento de um modelo baseado em GPT para o ensino de anamnese e documentação de prontuários médicos”, Revista Caderno Pedagógico, v. 22, n. 9, p. 1-15.

Jeevan, H. R. (2023) “The Evolution of Natural Language Processing”, Medium, https://medium.com/@jeevanchiru17/the-evolution-of-natural-language-processing-ad214cb9f6ca, October.

Jurafsky, D. and Martin, J. H. (2026) “Speech and Language Processing”, 3rd ed. draft, Stanford University.

Kamradt, G. (2024) “5 Levels of Text Splitting”, FullStackRetrieval Tutorials, https://github.com/FullStackRetrieval-com/RetrievalTutorials, April.

Karunanayake, N. (2025) “Next-generation agentic AI for transforming healthcare”, Informatics and Health, v. 2, p. 73-83. DOI: 10.1016/j.infoh.2025.03.001.

LangChain. (2026) “Retrieval”, LangChain Documentation, https://docs.langchain.com/v0.3/docs/concepts/retrieval, April.

Lewis, P. et al. (2020) “Retrieval-augmented generation for knowledge-intensive NLP tasks”, Advances in Neural Information Processing Systems (NeurIPS), v. 33, p. 9459–9474.

Li, J. et al. (2025) “Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents”, arXiv preprint arXiv:2405.02957v3.

Miquido. (2024) “What are guardrails in AI?”, https://www.miquido.com/ai-glossary/what-are-guardrails-in-ai/, April.

Müller, H. et al. (2022) “Explainability and causability for artificial intelligence-supported medical image analysis in the context of the European In Vitro Diagnostic Regulation”, New Biotechnology, v. 70, p. 67-72.

NVIDIA. (2024) “Programmable Guardrails: High-level flow”, NeMo Guardrails User Guide, v. 0.19.0, https://docs.nvidia.com/nemo/guardrails/0.19.0/user-guides/guardrails-process.html, April.

Nascimento, J. R. (2024) “Exploração de técnicas de engenharia de prompt para aprimorar os resultados do uso de LLM no TCMRio”, Trabalho de Conclusão de Curso (Especialização em TI), Instituto Metrópole Digital, UFRN, Natal.

Park, H. et al. (2025) “Scoping review of nurse triage in primary care”, BMC Nursing, v. 24, n. 1104. DOI: 10.1186/s12912-025-03740-3.

Radford, A. et al. (2018) “Improving language understanding by generative pre-training”, OpenAI, https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Rebedea, T. et al. (2023) “NeMo Guardrails: A toolkit for controllable and safe LLM applications with programmable rails”, arXiv preprint arXiv:2310.10501.

Suamchaiyaphum, K. et al. (2024) “Triage accuracy of emergency nurses: An evidence-based review”, Journal of Emergency Nursing, v. 50, n. 1, p. 44-54. DOI: 10.1016/j.jen.2023.10.001.

Tamkin, A. et al. (2021) “Understanding the capabilities, limitations, and societal impact of large language models”, arXiv preprint arXiv:2102.02503.

Touvron, H. et al. (2023) “LLaMA: Open and Efficient Foundation Language Models”, arXiv preprint arXiv:2302.13971.

van Kolfschooten, H. and van Oirschot, J. (2024) “The EU Artificial Intelligence Act (2024): Implications for healthcare”, Health Policy, v. 149, p. 105152.

Vaswani, A. et al. (2017) “Attention is all you need”, In: Advances in Neural Information Processing Systems, 30., Long Beach: NIPS, p. 5998-6008.

Wei, J. et al. (2022) “Chain-of-thought prompting elicits reasoning in large language models”, Advances in Neural Information Processing Systems, v. 35, p. 24824-24837.

Weidinger, L. et al. (2021) “Ethical and social risks of harm from language models”, DeepMind Research Report, arXiv preprint arXiv:2112.04359.

Zaniboni, J. V. N. (2025) “Integrando Técnicas de Geração Aumentada por Recuperação e Grandes Modelos de Linguagem para Auxílio ao Diagnóstico Médico”, Trabalho de Conclusão de Curso (Sistemas de Informação), UFSC, Florianópolis.

