2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a fundamentação teórica que sustenta a pesquisa, organizando o conhecimento em um percurso que parte do contexto do problema para chegar às soluções tecnológicas. A estrutura foi pensada para, primeiramente, delinear a complexidade do atendimento em IES brasileiras (Seção 2.1), estabelecendo os desafios práticos e de gestão que justificam a busca por inovação. Em seguida, o capítulo explora as experiências internacionais com o uso de atendentes virtuais no Ensino Superior (Seção 2.2), mapeando o estado da arte e as arquiteturas tecnológicas consolidadas. Essa progressão do "problema local" para as "soluções globais" permite contextualizar a contribuição desta pesquisa e fundamentar a arquitetura proposta neste trabalho.

2.1 O ATENDIMENTO NAS IES: COMPLEXIDADE, CONFIGURAÇÃO E CONTRADIÇÕES

O ponto de partida para a análise do atendimento em IES é o reconhecimento de sua natureza como organizações complexas. A universidade brasileira, em particular, pode ser compreendida através do modelo de "burocracia profissional" de Mintzberg (1995), no qual o poder operacional reside nos especialistas de cada unidade — docentes, pesquisadores e técnicos. Essa estrutura, constituída historicamente pela aglomeração de faculdades e departamentos semi-autônomos (Führ, 2022), promove uma cultura de fragmentação que dificulta a coordenação central e gera desafios críticos para a gestão da informação.

A consequência direta dessa autonomia departamental é a criação de silos de informação. O termo, uma metáfora da agricultura, descreve o fenômeno no qual departamentos dentro de uma organização se tornam isolados e com poucos meios de comunicação entre si (Cromity; De Stricker, 2011).

As barreiras que impedem o fluxo de dados não são apenas tecnológicas, mas fundamentalmente estratégicas e humanas (Miller; Tucker, 2014). Conforme diagnosticado por Carvalho *et al.* (2008, p. 9), as unidades organizacionais fecham-se sobre si mesmas, "criando silos que dificultam o fluxo inter-departamental. Cada grupo tem uma linguagem própria, uma sub-cultura e uma forma de se auto-organizar distinta dos demais".

Para o estudante, essa "balcanização" do conhecimento se traduz na necessidade de navegar por uma teia de setores para resolver uma única dúvida, materializando a complexidade já estudada em 1988:

> Muitas vezes a informação necessária à tomada de decisão ou à avaliação envolve dados de várias unidades, necessitando, dessa forma, ser coordenada, integrada e central. (Wolynec; Marin, 1988, p. 213)

É importante notar, contudo, que nem todo silo é indesejável; a proteção de dados sensíveis, por exemplo, exige isolamento legítimo. O problema central, e o foco desta pesquisa, são os "efeitos indesejados dos silos organizacionais que impedem o fluxo de informação" (Cromity; Stricker, 2011, p. 172), comprometendo a eficiência e a qualidade do atendimento.

Essa complexidade estrutural se projeta diretamente sobre a jornada do estudante. Forçado a interagir com múltiplos setores — da secretaria acadêmica à assistência estudantil — cada qual operando com normativas e sistemas próprios, o discente se vê diante de uma paisagem informacional fragmentada. Fontes de orientação estáticas, como FAQs e editais, frequentemente empregam um jargão técnico que eleva a barreira de acesso e gera frustração (Ramakrishnan *et al.*, 2024). Em resposta, os estudantes recorrem a canais de atendimento direto, como e-mail ou telefone, buscando informações que, embora publicadas, permanecem funcionalmente indisponíveis.

Esse comportamento, embora compreensível, alimenta um ciclo de retroalimentação da sobrecarga administrativa, onde servidores dedicam tempo precioso a demandas repetitivas. Este cenário, conforme apontam George e Wooden (2023), representa um obstáculo à transformação organizacional que exige a reavaliação de processos e o desenvolvimento de novas competências nos servidores.

Para compreender como superar esse obstáculo, é preciso, portanto, aprofundar a análise sobre a configuração do serviço que está no centro desse gargalo: o atendimento institucional. A expressão configuração do atendimento, neste contexto, refere-se ao modo como essa função se organiza no interior das universidades. O atendimento trata-se de uma atividade transversal, que conecta setores administrativos, acadêmicos, de pesquisa e de extensão, mediando a relação entre a instituição e sua comunidade.

A função transcende o mero "tira-dúvidas", abrangendo um espectro que vai de demandas objetivas (consulta a prazos, solicitação de documentos) a subjetivas (acolhimento, orientação psicossocial). Enquanto as primeiras demandam precisão e automação, as segundas exigem empatia e conhecimento tácito, aquele que, para ser processado por um sistema, precisa, antes de tudo, ser expresso por meio da linguagem humana (Duarte, 2003, p. 614). Demandas subjetivas são, por vezes, aquelas em que o profissional não se apoia sequer na fala do atendido para tomar a melhor decisão possível; pode ser um olhar, uma postura,

um suspiro. Essa distinção se torna ainda mais complexa quando consideramos a diversidade de usuários atendidos pelas instituições.

Considerando toda esta complexidade do ecossistema das IES, torna-se desafiador estabelecer uma definição única para o "atendimento". Ele pode manifestar-se na abertura de um chamado técnico sobre o sistema acadêmico, em uma dúvida sobre o empréstimo de livros, em uma ligação de familiares para tratar de documentos de matrícula ou, ainda, em um acolhimento social realizado por setores específicos.

Para os fins desta pesquisa, o atendimento é, portanto, enquadrado não como uma relação vertical de prestação de serviço, mas como um ato dialógico. Inspirando-se em Paulo Freire (1987), o diálogo autêntico é um "encontro de homens, mediatizados pelo mundo, para pronunciá-lo", uma relação horizontal de "eu-tu".

Transportando essa filosofia para o contexto institucional, o atendimento eficaz não é aquele em que a organização "deposita" informações no usuário, mas um encontro mediado pela necessidade informacional deste. Mesmo que a tecnologia seja a mediadora, a metodologia cíclica proposta na seção 3 foi concebida para que o sistema aprenda com a comunidade, em um processo que busca, idealmente, refletir essa relação de mutualidade.

2.2 REVISÃO BIBLIOGRÁFICA - IMPLANTAÇÕES E TECNOLOGIAS

A concepção de sistemas computacionais capazes de interagir em linguagem natural, embora idealizada desde os primórdios da inteligência artificial com o Teste de Turing (Turing, 1950) e experimentos seminais como o ELIZA (Weizenbaum, 1966), permaneceu por décadas restrita a arquiteturas procedimentais. Os sistemas pioneiros operavam a partir de regras e roteiros pré-definidos, limitando sua flexibilidade e capacidade de lidar com a complexidade de domínios não estruturados (Shum; He; Li, 2018).

A revolução impulsionada pela arquitetura Transformer Vaswani et al., 2017 e o subsequente advento dos LLMs inauguraram um novo paradigma, transformando os chatbots de meros executores de scripts em sistemas de raciocínio capazes de realizar tarefas complexas e fundamentar suas respostas em conhecimento externo.

Essa transição paradigmática abriu um campo fértil de exploração para as IES, cujo potencial para otimizar processos e ampliar o acesso à informação tem sido intensamente investigado. Compreender como essa evolução tecnológica foi aplicada e documentada no Ensino Superior é, portanto, um passo fundamental para contextualizar esta dissertação. Para mapear o estado da arte de forma sistemática, adotou-se o método construtivista de revisão

bibliográfica *Knowledge Development Process – Constructivist* (ProKnow-C), que oferece um processo rigoroso para a seleção de um portfólio de artigos alinhado aos interesses específicos do pesquisador (Ensslin *et al.*, 2010).

A construção do portfólio partiu de uma busca sistemática na base de dados Scopus, selecionada por sua relevância nas áreas de administração e tecnologia. A estratégia de busca foi desenhada para capturar a intersecção entre o contexto e a tecnologia, cruzando, por meio de operadores booleanos, descritores relativos ao ambiente de ensino com termos que denotam as tecnologias de linguagem investigadas.

A Tabela 2 apresenta os resultados quantitativos da busca inicial, revelando o panorama da produção científica na intersecção entre os eixos de pesquisa. Observa-se uma concentração significativa de publicações que associam os termos "Academic" e "University" a "LLM" e "Chatbot", somando mais da metade do total bruto de artigos, o que indica um expressivo interesse da comunidade científica na aplicação de modelos de linguagem e agentes conversacionais em contextos acadêmicos.

Em contrapartida, a baixa prevalência de artigos que combinam os termos com tecnologias mais específicas, como "Retrieval-Augmented Generation (RAG)" e "Semantic Search", sugere que, embora a aplicação geral de IA seja amplamente discutida, a investigação de arquiteturas mais sofisticadas ainda constitui um campo emergente e com menor documentação na literatura.

Tabela 2: Distribuição de Artigos Brutos por Cruzamento de Termos dos Eixos 1 e 2.

Pesquisa Booleana	Resultados
"Higher Education" AND ("LLM" OR "Language Model" OR "Large Language Model")	**249**
"Higher Education" AND ("Chatbot" OR "Chatbots" OR "Conversational Agent")	**295**
"Higher Education" AND ("Semantic Search" OR "Retrieval-Augmented Generation" OR "RAG")	**9**
"University" AND ("LLM" OR "Language Model" OR "Large Language Model")	**468**
"University" AND ("Chatbot" OR "Chatbots" OR "Conversational Agent")	**539**
"University" AND ("Semantic Search" OR "Retrieval-Augmented Generation" OR "RAG")	**67**
"Academic" AND ("LLM" OR "Language Model" OR "Large Language Model")	**786**
"Academic" AND ("Chatbot" OR "Chatbots" OR "Conversational Agent")	**546**
"Academic" AND ("Semantic Search" OR "Retrieval-Augmented Generation" OR "RAG")	**46**

TOTAL DE ARTIGOS CIENTÍFICOS	**3005**
TOTAL DE ARTIGOS CIENTÍFICOS SEM DUPLICATAS	**2189**

Fonte: Consulta em 03/05/2025 na base Scopus (2025).

Após a remoção de duplicatas, o processo resultou em um acervo de **2.189 artigos** únicos, confirmando a alta aderência e a sobreposição relevante das palavras-chave selecionadas. A etapa subsequente do ProKnow-C, que consiste na leitura e filtragem dos títulos para verificar o alinhamento com o tema específico desta dissertação, levou **à exclusão de 1.400 artigos** que, embora tangenciassem o assunto, não tratavam diretamente do uso de agentes conversacionais para atendimento em IES. Ao final desta fase, consolidou-se um portfólio preliminar de 789 publicações não repetidas e com alto potencial de relevância para a pesquisa.

A partir do **portfólio preliminar de 789 publicações**, o processo de seleção foi refinado. Em uma adaptação ao fluxo convencional do ProKnow-C, que sugere a aplicação imediata da análise de citações, optou-se primeiramente por uma filtragem qualitativa baseada na leitura dos resumos. Essa decisão metodológica visou garantir que apenas os artigos com real alinhamento temático fossem submetidos à análise de relevância científica. O critério de inclusão para esta etapa foi a aderência do resumo a pelo menos uma das seguintes condições:

a) tratar da aplicação de IA, LLMs ou agentes conversacionais no ambiente administrativo de IES; ou

b) apresentar uma contribuição tecnológica ou histórica relevante que, embora não focada no ensino superior, pudesse informar o design da arquitetura proposta.

A aplicação rigorosa desse filtro resultou na exclusão de 577 artigos, consolidando um banco de 212 publicações com alta aderência temática. Apenas neste ponto foi retomado o critério de reconhecimento científico do ProKnow-C. Utilizando dados de citação coletados na plataforma Google Scholar, verificou-se que os **212 artigos somavam 8.333 citações**. A análise da curva de Pareto revelou que **os 42 artigos mais citados (aproximadamente 20% do total) concentravam 5121 citações, representando 80% do impacto científico total do portfólio.**

A etapa final do processo consistiu na leitura integral dos 42 artigos de maior impacto científico, complementados por uma seleção estratégica de 4 publicações recentes. Embora excluídos pela curva de Pareto devido ao baixo número de citações, estes artigos

mais novos foram incluídos por seu alto potencial de alinhamento e por representarem o estado da arte mais atual. Essa leitura foi guiada por um duplo propósito:

1. extrair a fundamentação teórica para esta dissertação e, crucialmente;
2. mapear as experiências concretas de implantação de agentes conversacionais em IES.

O resultado final deste processo de seleção, sintetizado na Tabela 3, é um portfólio bibliográfico final de 21 artigos que constituem o núcleo desta revisão. Destes, **13 artigos se destacam por relatarem a implantação**, o teste e a avaliação de soluções de atendimento em instituições de ensino ao redor do mundo, fornecendo a base empírica para a análise que se segue.

Tabela 3 – Síntese do Processo de Revisão Bibliográfica (ProKnow-C)

Etapa	Quantidade de Artigos	Citações do Conjunto (Google Scholar)
Levantamento Bruto (sem duplicatas)	2.189	-
Filtro por Título	789	-
Filtro por Resumo	214	8.333
Seleção por Relevância Científica	42 (+4 recentes)	5.121 (+ 118 dos recentes)
Portfólio Bibliográfico (PB) Final	21	3.328
Recorte: Soluções Implantadas no PB	*12*	*1.290*

Fonte: Elaboração própria a partir da Base Scopus e registro de citações do Google Scholar(2025).

2.2.1 Estado da Arte: Aplicações Consolidadas de Agentes Virtuais no Ensino Superior

A partir do portfólio bibliográfico final, foram identificados 11 artigos que relatam a implantação e avaliação de agentes conversacionais no contexto do Ensino Superior. Para compreender o estado da arte e contextualizar o problema de pesquisa, estes casos foram sistematizados e classificados de acordo com sua arquitetura tecnológica.

A análise comparativa, apresentada nos quadros a seguir, revela uma clara trajetória evolutiva, partindo de sistemas baseados em regras e lógicas definidas, passando pela revolução dos LLMs no paradigma de agente único, e chegando a um movimento incipiente de modularização. Esta organização permite não apenas mapear as soluções existentes, mas

principalmente diagnosticar as limitações recorrentes que justificam a investigação de novas abordagens arquitetônicas.

2.2.1.1 Arquiteturas Fundamentais em IES Identificadas na Literatura

Quadro 1 - Panorama das Arquiteturas Fundamentais em IES

Estudo (Autor, Ano)	Chatbot, IES, País	Plataforma e Modelo de IA	Foco do Atendimento (Público; Problema)	Arquitetura
(Carayannopoulos, 2018)	BU111 Bot, Wilfrid Laurier Univ., Canadá	Plataforma de IM (Kik)	Calouros; mitigar sobrecarga informacional e apoiar transição	Baseada em Regras (fluxos/FAQ)
(Abbas *et al.*, 2022)	Bo, University of Leeds, Reino Unido	Plataforma de chat **Differ** (embutido)	**Estudantes maduros**; reduzir isolamento social e promover engajamento	**Não especificada** (interações contextuais; provável *script/flows*)*
(Al-Abdullatif; Al-Dokhny; Drwish, 2023)	Bashayer, King Faisal Univ., Arábia Saudita	Plataforma de IM (WhatsApp), Motor de Diálogo	**Pós-graduandos**; Aumentar a motivação e otimizar estratégias de aprendizagem.	Baseada em Recuperação (extrativa / *task-oriented*)
Nguyen et al. (2021)	NEU-chat bot, National Economics Univ., Vietnã	Rasa com BERT/DIET.	**Candidatos e pais**; Reduzir carga de trabalho do setor de admissões.	Baseada em Recuperação (extrativa)
Ramakrishnan *et al.*, 2024)	Pacific Chatbot, Univ. of the Pacific, EUA	Customizado, Rede Neural	**Candidatos/Ingressantes**; Otimizar a busca por FAQs e reduzir a carga administrativa.	Classificação supervisionada (intents/FAQ)
(Chen *et al.*, 2023)	Sammy, San Jose State University, EUA	Plataforma No-code (Juji)	**Estudantes de Negócios**; Ensinar conceitos de IA e levantar requisitos de atendimento.	Híbrida (Regras + Deep Learning)

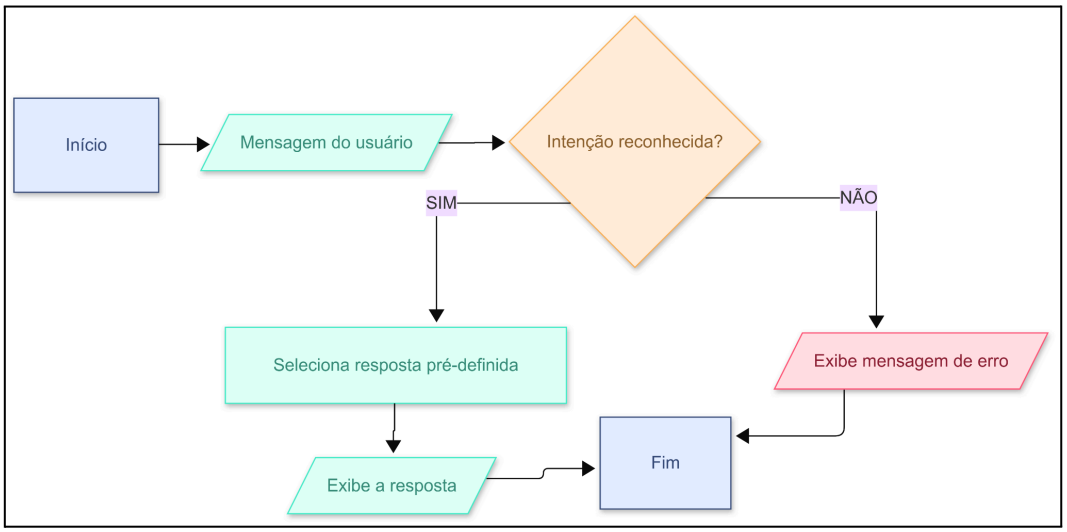
Fonte: Elaboração Própria (2025)

A análise do Quadro 1 revela que as abordagens fundacionais de atendimento em IES, mesmo em estudos recentes, frequentemente se baseiam em fluxos de diálogo predefinidos e lógicas determinísticas. O BU111 Bot (Carayannopoulos, 2018) e o chatbot Bo (Abbas *et al.*, 2022) exemplificam essa categoria. Ambos foram projetados para apoiar a integração de públicos específicos por meio de interações estruturadas, como ice breakers e respostas a FAQs.

Esses casos demonstram que, mesmo com flexibilidade limitada, os chatbots podem transcender a recuperação de informação para atuar na promoção do engajamento e do senso de pertencimento. A arquitetura subjacente a esses sistemas pioneiros é a baseada em regras, que representa a abordagem fundacional dos agentes conversacionais. Conhecidos também como chatbots estruturados, eles operam a partir de fluxos de conversação predeterminados, seguindo uma lógica inerentemente determinística de "se-então" (Villegas-Ch *et al.*, 2021).

Seu funcionamento se apoia em regras explícitas: ao receber uma mensagem, o sistema utiliza técnicas como correspondência de padrões para identificar a intenção e selecionar uma resposta de um repertório pré-escrito. Sua incapacidade de lidar com consultas não previstas no roteiro é uma de suas principais características (Janssen *et al.*, 2020), como ilustra o fluxo na Figura 2.

Figura 2: Fluxo de Funcionamento de um Chatbot Baseado em Regras



Fonte: Elaboração Própria (2025).

A simplicidade estrutural desses sistemas é sua maior vantagem, tornando-os viáveis para educadores sem conhecimento técnico aprofundado (Hew *et al.*, 2023). Contudo, essa mesma simplicidade limita sua eficácia frente a consultas complexas ou ambíguas, podendo

gerar frustração no usuário e exigindo manutenção constante para manter as respostas atualizadas (Lappalainen; Narayanan, 2023).

Para superar essa rigidez, a etapa evolutiva seguinte introduz mecanismos de inteligência artificial. O Bashayer Chatbot (Al-Abdullatif *et al.*, 2023) emprega uma arquitetura baseada em recuperação (retrieval-based), buscando respostas em um repositório de interações para guiar estudantes. O NEU-chatbot (Nguyen et al., 2021) reforça esse degrau intermediário: uma recuperação extrativa com Rasa e BERT/DIET, voltada a admissões. Em domínio restrito e repetitivo, reporta alta acurácia e alívio operacional, evidência de que especialização de escopo potencializa desempenho mesmo sem geração por LLM.

O Sammy Chatbot (CHEN *et al.*, 2023) avança para um modelo híbrido, combinando regras com deep learning. Embora os estudantes tenham avaliado positivamente sua usabilidade e interatividade, a análise qualitativa revelou as limitações da inteligência centralizada. Os participantes descreveram o diálogo como 'transacional' e criticaram a incapacidade do sistema de compreender nuances, como gírias ou erros de digitação, e a falta de uma 'conexão emocional'. Esses desafios demonstram que, mesmo em arquiteturas híbridas, o agente único luta para simular a fluidez e a profundidade de uma interação humana genuína.

O estudo do Pacific Chatbot (Ramakrishnan *et al.*, 2024) aprofunda essa discussão por meio de uma análise comparativa explícita. Ao testar um modelo de propósito geral (BERT) contra uma rede neural customizada para a classificação de FAQs, os resultados foram contundentes: a abordagem especializada alcançou uma precisão superior (98% vs. 87%). Os autores observam que a performance inferior do BERT se deu justamente porque o dataset era "pequeno e customizado" (Ramakrishnan *et al.*, 2024, p.47), evidenciando os limites de modelos generalistas em contextos de alta especificidade.

Adicionalmente, o estudo expõe os desafios operacionais inerentes à manutenção de qualquer sistema baseado em regras: a necessidade de limpeza manual de dados para lidar com informações desatualizadas; a dependência de treinamento regular para incorporar novas FAQs; a ausência de um sistema de feedback automatizado, apontada como uma limitação para a evolução do chatbot.

Em conjunto, esses casos mostram a trilha: regras → recuperação extrativa → híbridos → classificação supervisionada, ainda sob o paradigma de agente único. A próxima seção examina como os LLMs introduzem RAG e ampliam o teto de capacidade, ao custo de novos desafios de contexto e alucinação.

A próxima seção analisa como a introdução dos Modelos de Linguagem de Grande Escala (LLMs) redefiniu o potencial desses agentes, ao mesmo tempo que expôs novas fronteiras e desafios.

2.2.1.2 Arquiteturas RAG de Agente Único em IES identificadas na literatura

O advento dos LLMs representa o segundo grande paradigma arquitetônico. Diferentemente dos sistemas baseados em regras, os chatbots generativos operam sobre uma hierarquia tecnológica sofisticada. Sua fundação é o Processamento de Linguagem Natural (PNL), campo que dota as máquinas da capacidade de interpretar e gerar linguagem humana (Nguyen *et al.*, 2021). Os LLMs são o motor avançado que executa as tarefas de PNL, antecipando palavras em uma sequência para construir textos coerentes e contextualmente relevantes (Lappalainen; Narayanan, 2023).

A evolução dos chatbots generativos, contudo, expôs uma limitação crítica para aplicações institucionais: a ausência de um mecanismo intrínseco de verificação factual, resultando no fenômeno das alucinações (Salemi; Zamani, 2024). Em um ambiente como o de uma IES, onde a precisão de informações sobre normas e currículos é inegociável, a confiança no sistema torna-se um requisito fundamental (Bilquise; Ibrahim; Salhieh, 2024).

A evolução dos chatbots generativos, embora represente um salto em fluidez conversacional, expôs uma limitação crítica para aplicações institucionais: a ausência de um mecanismo intrínseco de verificação factual, resultando no fenômeno das alucinações (Salemi; Zamani, 2024). Em um ambiente complexo como o de uma IES, onde a precisão de informações sobre prazos, normas e currículos é inegociável, a confiança na resposta do sistema torna-se um requisito fundamental (Bilquise; Ibrahim; Salhieh, 2024).

Para superar esse desafio, emerge um novo paradigma arquitetural que transcende os chatbots tradicionais: o *Agentic RAG*. Essa abordagem combina dois componentes essenciais:

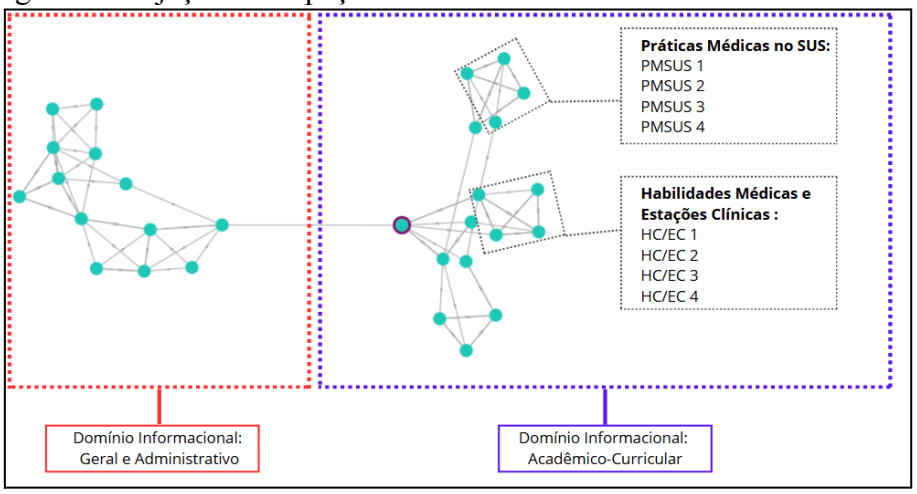
1. **O agente como núcleo de raciocínio:** diferentemente de um chatbot que apenas gera texto, um agente opera sobre um LLM para raciocinar, planejar e, sobretudo, utilizar ferramentas externas (Wang *et al.*, 2024) . Ele não está limitado ao conhecimento adquirido durante o pré-treinamento; é capaz de decidir ativamente buscar informações em APIs, bancos de dados ou outros sistemas externos para cumprir uma tarefa específica (Wu et al., 2025).

2. **O RAG**, proposto originalmente por Lewis *et al.*, 2021, como ferramenta de fundamentação: atua como principal mecanismo de "consulta à banco de dados" para o agente.

A eficácia do Agentic RAG depende inteiramente da qualidade da preparação dos dados. A fase de indexação inicia-se com a organização de documentos institucionais em blocos de informação, os chunks, que são então transformados em vetores numéricos (embeddings). Este processo de vetorização, pode ser compreendido visualmente através de um exemplo prático desenvolvido nesta pesquisa.

A Figura 3 apresenta a projeção bidimensional do espaço de *embeddings* de um protótipo que responde sobre o curso de um determinado curso de Medicina, demonstrando como a teoria da similaridade semântica se traduz em organização espacial. Na visualização, chunks com conteúdo semântico similar são agrupados em clusters distintos: trechos sobre processos administrativos, por exemplo, formam um agrupamento coeso, geometricamente distante do cluster que reúne informações acadêmico-curriculares.

Figura 3: Projeção do Espaço Vetorial Evidenciando Clusters Semânticos



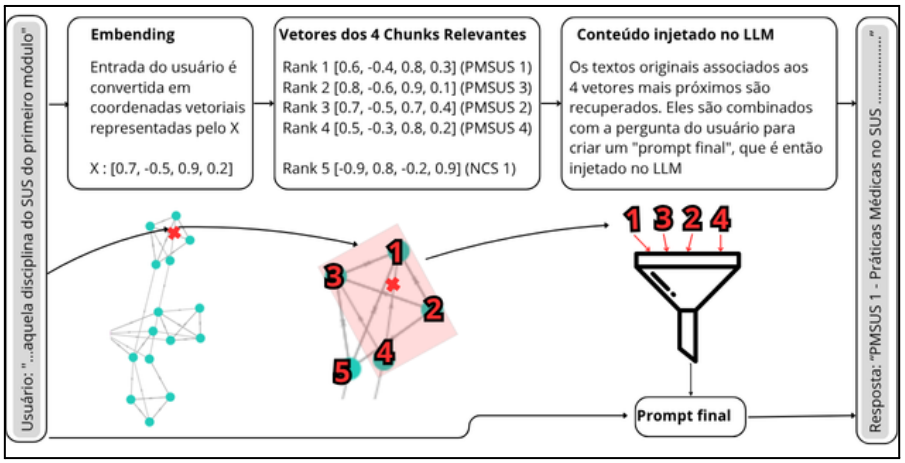
Fonte: Elaboração Própria (2025).

O ponto mais relevante desta análise reside na estrutura interna do *cluster* Acadêmico-Curricular. Mesmo quando os chunks compartilham um domínio informacional comum, o modelo de embedding é capaz de criar subagrupamentos ainda mais refinados, como os clusters para os eixos disciplinares de "Práticas Médicas no SUS" e "Habilidades Médicas".

Compreendida a preparação da base de conhecimento, o ciclo de funcionamento de recuperação em um sistema RAG, ilustrado na Figura 4, inicia-se com a conversão da pergunta do usuário em outro vetor numérico, ou *embedding* (1), que define sua posição no

espaço semântico. Este vetor é então utilizado para consultar a base de conhecimento e recuperar os chunks de informação textualmente mais próximos neste espaço (2).

Figura 4: Recuperação de Informação e Geração de Texto em um Sistema RAG.



Fonte: Elaboração Própria adaptada de (Wu et al., 2025, p. 3).

Os textos originais desses chunks são combinados com a pergunta para criar um prompt final robusto (3), que é então enviado ao LLM para a geração da resposta (4). Este mecanismo funciona como uma "memória externa" que obriga o modelo a basear sua resposta no conteúdo factual fornecido, elevando significativamente a precisão e a confiabilidade das interações (WU et al., 2025).

Compreendido esse mecanismo, a Figura 5, a seguir, demonstra o resultado final deste processo em uma interação simulada. Nela, um usuário busca informações de forma coloquial ("aquela disciplina sobre o SUS") e o agente, fundamentado pelo RAG, é capaz de identificar o chunk mais relevante e fornecer uma resposta precisa e contextualizada.