INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS PIRIPIRI
TECNÓLOGO EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

LARISSA SOUZA DO NASCIMENTO

DESENVOLVIMENTO DE UMA SOLUÇÃO TECNOLÓGICA UTILIZANDO APIS DE INTELIGÊNCIA ARTIFICIAL GENERATIVA PARA OTIMIZAÇÃO DE PROCESSOS INSTITUCIONAIS DO IFPI

PIRIPIRI
2026

LARISSA SOUZA DO NASCIMENTO

DESENVOLVIMENTO DE UMA SOLUÇÃO TECNOLÓGICA UTILIZANDO APIS DE INTELIGÊNCIA ARTIFICIAL GENERATIVA PARA OTIMIZAÇÃO DE PROCESSOS INSTITUCIONAIS DO IFPI

Trabalho de Conclusão de Curso (artigo científico) apresentado como exigência parcial para obtenção do diploma do Curso de Tecnólogo em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Piripiri.

Orientador: Prof. Dr. Ricardo Moura Sekeff Budaruiche

PIRIPIRI
2026

Desenvolvimento de uma Solução Tecnológica Utilizando APIs de Inteligência Artificial Generativa para Otimização de Processos Institucionais do IFPI

Larissa Souza do Nascimento[1]
Prof. Dr. Ricardo Moura Sekeff Budaruiche[2]

RESUMO

Instituições de ensino mantêm grande volume de documentos normativos e administrativos, cujo acesso é dificultado pela dispersão, atualização frequente e complexidade da linguagem. Este trabalho apresenta uma solução conversacional inteligente para consulta de informações do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, integrando inteligência artificial generativa, geração aumentada por recuperação, sistemas multiagentes e um mecanismo de atualização contínua da base de conhecimento inspirado na teoria de revisão de crenças. A arquitetura emprega agentes especializados em recuperação de informações, processamento documental e manutenção da consistência do conhecimento. A avaliação combinou testes técnicos automatizados e participação de usuários. Em 51 questões institucionais, o sistema obteve taxa de execução de 100%, acurácia de fundamentação em fontes institucionais de 78%, similaridade semântica média de 0,78, acurácia semântica de 63% e média de sobreposição de 0,54 entre fontes recuperadas e esperadas. O mecanismo de revisão de crenças apresentou comportamento consistente em 65 cenários de verificação com documentos sintéticos e institucionais reais. Na avaliação com 34 participantes, a solução alcançou média de 70,8 no Questionário de Usabilidade de Chatbots e nota geral de 8,1 em uma escala de 0 a 10. Os resultados indicam que a abordagem é viável para apoiar o acesso a informações institucionais, a atualização contínua do conhecimento, a confiabilidade das respostas e a experiência dos usuários.

Palavras-chave: inteligência artificial generativa; revisão de crenças; sistemas multiagentes; *chatbot* institucional; geração aumentada por recuperação.

ABSTRACT

Educational institutions maintain a large volume of normative and administrative documents whose access is hindered by dispersion, frequent updates, and complex language. This work presents an intelligent conversational solution for consulting information from the Federal Institute of Education, Science and Technology of Piauí, integrating generative artificial intelligence, retrieval-augmented generation, multi-agent systems, and a continuous knowledge base update mechanism inspired by belief revision theory. The architecture employs specialized agents for information retrieval, document processing, and knowledge consistency maintenance. The evaluation combined automated technical tests and user participation. Across 51 institutional questions, the system achieved a 100% execution rate, 78% grounding accuracy in institutional sources, an average semantic similarity of 0.78, a semantic accuracy of 63%, and an average

[1]Discente do curso de Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Piripiri. E-mail: larysou66@gmail.com.

[2]Doutor em Ciência da Computação, pesquisador e profissional de tecnologia com atuação em inteligência artificial, sistemas de informação, análise de dados e ontologias. Docente do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Piripiri. E-mail: ricardo.sekeff@ifpi.edu.br.

overlap of 0.54 between retrieved and expected sources. The belief revision mechanism demonstrated consistent behavior in 65 verification scenarios involving synthetic and real institutional documents. In the evaluation with 34 participants, the solution achieved an average score of 70.8 on the Chatbot Usability Questionnaire and an overall rating of 8.1 on a scale from 0 to 10. The results indicate that the approach is viable for supporting access to institutional information, continuous knowledge updating, response reliability, and user experience.

Keywords: generative artificial intelligence; belief revision; multi-agent systems; institutional chatbot; retrieval-augmented generation.

Data de aprovação: 14/07/2026

1 INTRODUÇÃO

A crescente digitalização dos processos acadêmicos e administrativos em instituições públicas de ensino tem intensificado a produção e a dependência de documentos normativos, como resoluções, portarias e manuais operacionais. No contexto do Instituto Federal de Educação, Ciência e Tecnologia do Piauí (IFPI), esses documentos desempenham papel fundamental na regulação das atividades institucionais, orientando tanto o corpo docente e técnico-administrativo quanto os discentes na execução de procedimentos acadêmicos e administrativos. No entanto, o elevado volume, a dispersão das fontes e a complexidade da linguagem normativa tornam o acesso e a interpretação dessas informações um desafio recorrente, impactando diretamente a eficiência dos processos e a qualidade do atendimento institucional.

Nesse contexto, soluções baseadas em Inteligência Artificial (IA), especialmente aquelas fundamentadas em modelos de linguagem de grande porte (*Large Language Models* – LLMs), têm se mostrado promissoras para automatizar a recuperação e interpretação de informações. Abordagens como *Retrieval-Augmented Generation* (RAG) permitem que sistemas conversacionais utilizem bases documentais institucionais para fornecer respostas mais precisas e contextualizadas. Entretanto, tais abordagens apresentam limitações importantes, principalmente no que se refere à manutenção da consistência do conhecimento ao longo do tempo, considerando que documentos normativos estão sujeitos a atualizações, revogações e possíveis contradições. Embora existam diversos trabalhos que empreguem RAG em sistemas conversacionais, ainda são pouco exploradas soluções voltadas à atualização contínua da base de conhecimento diante de mudanças em documentos normativos institucionais, evidenciando uma lacuna que motiva esta pesquisa.

Delimita-se, assim, o problema de pesquisa que orienta este trabalho: sistemas conversacionais institucionais baseados em RAG não dispõem, em sua formulação convencional, de mecanismos que identifiquem alterações nos documentos normativos indexados e ajustem a base de conhecimento em consequência, de modo que

respostas fundamentadas em versões superadas dessas normativas permanecem indistinguíveis, para o usuário, daquelas fundamentadas em versões vigentes. A partir dessa delimitação, formula-se a seguinte pergunta central: de que modo um mecanismo inspirado na teoria de Revisão de Crenças pode ser incorporado a uma arquitetura conversacional baseada em RAG e Sistemas Multiagentes (SMA), de maneira a identificar alterações em documentos normativos institucionais e refletir tais alterações na base de conhecimento, e que resultados essa incorporação apresenta quanto à fundamentação das respostas e à usabilidade percebida pelos usuários?

Para responder a essa pergunta, define-se como objetivo geral desenvolver e avaliar uma solução conversacional voltada à consulta de informações institucionais do IFPI que integre IA generativa, RAG e SMA a um mecanismo de atualização contínua da base de conhecimento inspirado na teoria de Revisão de Crenças. Como objetivos específicos, estabelecem-se: (i) projetar uma arquitetura capaz de recuperar informações institucionais de forma confiável; (ii) implementar um mecanismo de atualização contínua da base de conhecimento que identifique alterações em documentos normativos; e (iii) avaliar a eficácia da solução por meio de experimentos técnicos e da percepção dos usuários quanto à qualidade das respostas e à experiência de uso.

Em atendimento a esses objetivos, propõe-se o desenvolvimento de um *chatbot* institucional inteligente que integra técnicas de IA generativa, recuperação de informação por meio de RAG, arquiteturas baseadas em SMA e um mecanismo inspirado na teoria de Revisão de Crenças para promover a atualização contínua da base de conhecimento. Diferentemente de abordagens tradicionais, o sistema é capaz de identificar alterações semânticas em documentos institucionais, detectar obsolescência de informações e decidir automaticamente pela manutenção, atualização ou sinalização de obsolescência de conteúdos previamente armazenados.

A solução adota uma arquitetura baseada em Sistemas Multiagentes, utilizando o *framework* de orquestração agêntica LangGraph, no qual diferentes agentes especializados atuam de forma coordenada para realizar tarefas como busca, extração, processamento, validação e armazenamento de informações. Um agente de revisão de crenças atua como núcleo decisório do sistema, avaliando continuamente a consistência das informações com base no critério de similaridade de cosseno, integridade das fontes e variações estruturais nos documentos. Essa abordagem favorece maior modularidade, escalabilidade e robustez na gestão do conhecimento institucional. Para isso, a arquitetura organiza o fluxo de processamento em duas frentes principais:

1. *pipeline* de resposta ao usuário (Figura 2), no qual um orquestrador inteligente direciona as consultas para agentes de busca institucional ou externa;

2. *pipeline* de atualização contínua da base de conhecimento, responsável pela ingestão e validação de documentos institucionais, demonstrado na Figura 3.

Esse modelo híbrido possibilita a recuperação eficiente de informações e busca favorecer o alinhamento das respostas às versões mais recentes das normativas institucionais. Ressalta-se, contudo, que tal alinhamento não é assegurado de forma absoluta no protótipo desenvolvido, pois, conforme detalhado na Seção 4.2.5, documentos confirmados como obsoletos permanecem na base vetorial e ainda podem ser recuperados, uma vez que a filtragem por metadados durante a busca semântica não foi implementada nesta versão. Dessa forma, o trabalho integra IA generativa, orquestração multiagente e mecanismos de atualização contínua do conhecimento, buscando ampliar o acesso à informação e preservar sua consistência ao longo do tempo.

A validação da proposta em relação aos objetivos estabelecidos é conduzida em duas frentes complementares, descritas na Seção 3.1 e discutidas na Seção 5: uma avaliação técnica automatizada, voltada à fundamentação das respostas e ao comportamento do mecanismo de revisão de crenças, e uma avaliação com usuários da comunidade acadêmica, voltada à usabilidade percebida. A pergunta central é retomada na Seção 6, à luz dos resultados obtidos em ambas as frentes. A solução proposta apresenta potencial para contribuir com a redução da sobrecarga do corpo técnico-administrativo e com o aumento da autonomia dos usuários, benefícios que poderão ser ampliados à medida que sistemas dessa natureza forem incorporados às rotinas institucionais.

O artigo está estruturado da seguinte forma: a Seção 2 apresenta o referencial teórico que fundamenta a pesquisa, abordando conceitos relacionados à Inteligência Artificial Generativa, *Retrieval-Augmented Generation*, Sistemas Multiagentes e Revisão de Crenças. A Seção 3 descreve a metodologia adotada, incluindo a caracterização da pesquisa e os procedimentos de avaliação empregados. A Seção 4 apresenta a solução proposta, detalhando a arquitetura multiagente, o *pipeline* de atualização e revisão de crenças, as tecnologias utilizadas e as estratégias de segurança. A Seção 5 apresenta e discute os resultados obtidos nas avaliações técnicas e com usuários. Por fim, a Seção 6 expõe as conclusões do trabalho, suas limitações e as perspectivas para pesquisas futuras.

2 REFERENCIAL TEÓRICO

Esta seção apresenta os fundamentos teóricos que sustentam o desenvolvimento da solução proposta, estabelecendo a relação entre os conceitos da literatura e as decisões arquiteturais adotadas no sistema.

2.1 INTELIGÊNCIA ARTIFICIAL GENERATIVA E MODELOS DE LINGUAGEM

A Inteligência Artificial generativa tem se destacado como uma das principais abordagens para o processamento e a geração de linguagem natural, sobretudo com

o avanço dos modelos fundamentados na arquitetura *Transformer*, introduzida por Vaswani *et al.* (2017). Tais modelos demonstram capacidade de capturar relações semânticas complexas em textos, viabilizando a geração de respostas coerentes e contextualizadas a partir de entradas em linguagem natural. Nesse contexto, os LLMs têm sido amplamente empregados em sistemas conversacionais, possibilitando a automação de tarefas como atendimento ao usuário, interpretação de documentos e suporte à tomada de decisão. Conforme apontado por Zhao *et al.* (2023), esses modelos apresentam capacidades notáveis, mas também limitações relevantes, entre as quais se destacam a geração de informações imprecisas, fenômeno denominado "alucinações", e a ausência de conhecimento atualizado acerca de domínios específicos.

Em ambientes institucionais, tais limitações tornam-se ainda mais críticas, uma vez que respostas incorretas podem impactar diretamente processos administrativos e acadêmicos. Nessa perspectiva, Saad e Qawaqneh (2024) propõem o BrockportGPT, um *chatbot* institucional de domínio fechado voltado ao contexto acadêmico. Os autores comparam diferentes abordagens de *question answering*, incluindo treinamento do zero, *fine-tuning* e RAG, concluindo que a estratégia baseada em recuperação de informação apresenta maior precisão e reduz significativamente a ocorrência de "alucinações", ao restringir as respostas a fontes institucionais confiáveis. Diante disso, evidencia-se a necessidade de complementar o uso de LLMs com mecanismos que assegurem maior controle sobre as fontes de informação consultadas, especialmente em cenários nos quais a confiabilidade e a atualidade dos dados constituem requisitos essenciais.

2.2 *RETRIEVAL-AUGMENTED GENERATION* (RAG)

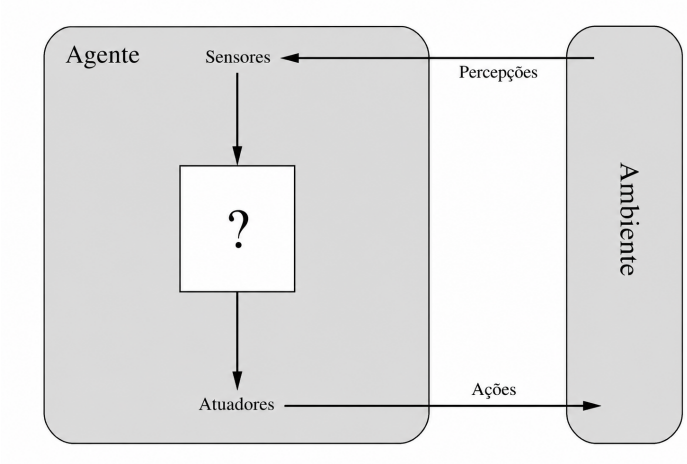
Como forma de mitigar as limitações dos modelos generativos, a abordagem de *Retrieval-Augmented Generation*, proposta por Lewis *et al.* (2020), tem sido amplamente adotada em sistemas que demandam maior precisão informacional. A técnica combina a capacidade de geração de linguagem dos LLMs com mecanismos de recuperação de informação, permitindo que o modelo fundamente suas respostas em documentos previamente indexados. Nesse contexto, os documentos são convertidos em representações numéricas denominadas *embeddings*, que consistem em vetores capazes de capturar características semânticas dos textos, possibilitando que conteúdos com significados semelhantes sejam posicionados próximos em um espaço vetorial. A partir de uma consulta do usuário, o sistema recupera os documentos mais relevantes por meio da comparação entre esses vetores. Métodos de recuperação densa, como os apresentados por Karpukhin *et al.* (2020), possibilitam a busca eficiente por similaridade semântica em grandes coleções documentais, sendo os documentos recuperados empregados como contexto para a geração da resposta, o que aumenta a confiabilidade e reduz a ocorrência de "alucinações".

Adicionalmente, a abordagem RAG permite separar o conhecimento paramétrico, armazenado nos pesos do modelo, do conhecimento externo, mantido em bases de dados estruturadas, conferindo maior controle, interpretabilidade e facilidade de atualização das informações mobilizadas durante a geração de respostas. Apesar dessas vantagens, o RAG apresenta desafios inerentes à manutenção da base de conhecimento. Em cenários nos quais os documentos estão sujeitos a alterações frequentes, como ocorre com normativas e regulamentos institucionais, faz-se necessário um mecanismo capaz de identificar mudanças e assegurar que a base vetorial reflita o estado mais recente da informação disponível. Essa limitação motiva a integração de mecanismos adicionais de gestão do conhecimento, como a abordagem de revisão de crenças adotada no presente trabalho.

2.3 SISTEMAS MULTIAGENTES E ORQUESTRAÇÃO DE IA

A complexidade envolvida na integração de múltiplas etapas de processamento, como coleta de dados, extração, análise semântica e geração de respostas, torna necessária a adoção de arquiteturas modulares. Nesse contexto, destacam-se os Sistemas Multiagentes, nos quais diferentes agentes autônomos colaboram para a execução de tarefas específicas. Segundo Russell e Norvig (2021, p. 34), “um agente é tudo o que pode ser considerado capaz de perceber seu ambiente por meio de sensores e de agir sobre esse ambiente por intermédio de atuadores”. Essa concepção, ilustrada na Figura 1, fornece a base teórica para o desenvolvimento de sistemas multiagentes, amplamente empregados na decomposição de problemas complexos em componentes especializados e cooperativos.

Figura 1 – Agente interagindo com o ambiente por meio de sensores e atuadores.



Fonte: Russell e Norvig (2021, p. 34).

Cada agente é responsável por uma função bem definida, como recuperação de dados, processamento de linguagem ou validação de informações, permitindo maior

especialização e eficiência no sistema como um todo. A coordenação entre esses agentes é realizada por um orquestrador, responsável por gerenciar o fluxo de execução e tomar decisões com base no contexto da requisição. Essa organização favorece a modularidade da solução, possibilitando a separação de responsabilidades entre componentes especializados e facilitando a manutenção, a escalabilidade e a adaptação do sistema a novas demandas.

A abordagem apresenta vantagens como modularidade, escalabilidade, facilidade de manutenção e possibilidade de evolução incremental. Com o avanço recente dos LLMs, sistemas multiagentes passaram a ser empregados em conjunto com modelos de linguagem para a construção de arquiteturas mais sofisticadas, nas quais agentes especializados interagem entre si para a resolução de tarefas complexas (Zhu *et al.*, 2026). Essa integração viabiliza a distribuição de responsabilidades entre diferentes componentes, reduzindo a sobrecarga imposta a um único modelo e aumentando a confiabilidade das respostas geradas. Na arquitetura proposta, a organização multiagente permite separar de forma clara os fluxos de atualização do conhecimento e de resposta ao usuário, além de possibilitar a inclusão de um agente especializado na revisão de crenças.

2.4 REVISÃO DE CRENÇAS E ATUALIZAÇÃO DE CONHECIMENTO

A manutenção de bases de conhecimento consistentes e atualizadas representa um dos principais desafios da área de representação do conhecimento, especialmente em cenários nos quais novas informações são incorporadas continuamente e podem entrar em conflito com conteúdos previamente armazenados. Nesse contexto, destaca-se a teoria de revisão de crenças AGM (Alchourrón, Gärdenfors e Makinson), proposta por Alchourrón, Gärdenfors e Makinson (1985), que estabelece princípios formais para a atualização racional de bases de conhecimento. Seu objetivo é definir como um agente deve modificar suas crenças diante de novas informações, preservando ao máximo o conhecimento já existente e garantindo a consistência do conjunto resultante.

A teoria AGM define três operações fundamentais para a modificação de crenças: *expansão*, *revisão* e *contração*. A operação de **expansão** consiste na adição de uma nova informação à base de conhecimento sem a remoção de conteúdos previamente existentes. A **revisão** é utilizada quando a nova informação pode gerar inconsistências, exigindo alterações no conjunto de crenças para restaurar sua coerência lógica. Já a **contração** corresponde à remoção de uma crença anteriormente aceita, permitindo descartar informações que deixaram de ser válidas. Essas operações constituem a base conceitual de diversos mecanismos de atualização de conhecimento empregados em sistemas inteligentes.

Embora a teoria AGM tenha sido originalmente concebida para bases de conhe-

cimento representadas por meio de lógica formal, seus princípios podem ser adaptados para contextos que utilizam documentos textuais e técnicas modernas de Processamento de Linguagem Natural (PLN). Em sistemas baseados em arquiteturas RAG, o conhecimento é armazenado em textos não estruturados e recuperado por meio de representações vetoriais e mecanismos de busca semântica. Nesses casos, a aplicação direta dos postulados formais da teoria torna-se inviável, uma vez que as informações não estão organizadas em proposições lógicas explícitas, mas em representações distribuídas que capturam relações de significado entre os conteúdos.

Diante dessa limitação, este trabalho utiliza os princípios da revisão de crenças como fundamento conceitual para o mecanismo de atualização da base de conhecimento do *chatbot*. A abordagem proposta emprega *embeddings* e medidas de similaridade semântica por cosseno para comparar diferentes versões de documentos institucionais e identificar alterações relevantes em seu conteúdo. A similaridade de cosseno consiste em uma métrica amplamente utilizada em recuperação de informação para avaliar o grau de proximidade entre vetores, considerando o ângulo formado entre eles (Manning; Raghavan; Schütze, 2008). Dessa forma, documentos ou trechos textuais semanticamente semelhantes tendem a apresentar valores mais elevados de similaridade, mesmo quando não compartilham exatamente os mesmos termos. Com base nessa análise, o sistema é capaz de detectar informações inseridas, modificadas ou removidas, atualizando apenas os trechos afetados da base vetorial e evitando o reprocessamento completo dos documentos.

Assim, operações análogas às definidas pela teoria AGM são realizadas de forma compatível com a representação textual adotada pelo sistema. Nesse contexto, a incorporação de novos conteúdos pode ser associada à operação de expansão, enquanto a atualização de informações previamente existentes aproxima-se do conceito de revisão. De maneira semelhante, a inatividade de trechos considerados obsoletos corresponde à operação de contração. Essa adaptação permite manter a coerência e a atualidade da base de conhecimento utilizada pelo *chatbot*, contribuindo para que as respostas geradas reflitam as versões mais recentes das normativas institucionais e reduzindo o impacto de inconsistências decorrentes de alterações documentais.

3 METODOLOGIA

O trabalho caracteriza-se como uma pesquisa aplicada, de natureza exploratória, com abordagem qualitativa-quantitativa, cujo objetivo é projetar, implementar e avaliar um sistema conversacional inteligente voltado à interpretação de normativas institucionais do IFPI. A proposta fundamenta-se na integração de técnicas de Inteligência Artificial generativa com uma arquitetura baseada em agentes inteligentes, incorporando mecanismos de atualização contínua do conhecimento inspirados no conceito de revisão de crenças.

A condução da pesquisa foi estruturada em cinco eixos principais: (i) modelagem da arquitetura do sistema; (ii) implementação do mecanismo de revisão de crenças; (iii) seleção e integração das tecnologias empregadas; (iv) implementação da camada de segurança contra *prompt injection*; e (v) avaliação experimental da solução. Os quatro primeiros eixos correspondem às etapas de construção do artefato, cujo resultado é apresentado na Seção 4; o quinto, relativo aos procedimentos de avaliação, é detalhado nas subseções seguintes.

3.1 ESTRATÉGIAS DE AVALIAÇÃO

A avaliação da solução proposta foi estruturada de forma multimodal, combinando diferentes estratégias complementares com o objetivo de analisar tanto o desempenho técnico do sistema quanto a percepção dos usuários em contexto real de uso. Essa abordagem busca superar limitações de métodos isolados, integrando métricas quantitativas automatizadas com avaliação qualitativa baseada na experiência do usuário.

3.1.1 Avaliação Técnica Automatizada

A avaliação do sistema foi conduzida em três frentes complementares: (i) avaliação técnica das respostas geradas pelo *chatbot*, (ii) verificação do mecanismo de revisão de crenças e (iii) testes de integração e consistência da arquitetura.

A avaliação das respostas foi realizada por meio de uma estratégia *offline*, baseada na comparação entre respostas produzidas pelo sistema e respostas de referência previamente definidas. Para isso, foi construído um conjunto de 51 questões representativas do domínio institucional, contemplando regulamentos acadêmicos, procedimentos administrativos, calendários, horários de cursos e informações funcionais extraídas dos documentos indexados na base de conhecimento. Cada item do conjunto de avaliação foi composto por uma pergunta, uma resposta esperada e uma ou mais fontes de referência. A avaliação das respostas foi realizada por meio das seguintes métricas:

- *Semantic Accuracy*: proporção de respostas consideradas semanticamente corretas em relação ao gabarito, isto é, aquelas cuja similaridade semântica supera um limiar previamente estabelecido, independentemente de diferenças na redação.
- *Avg Semantic Similarity*: média da similaridade semântica entre as respostas geradas pelo sistema e as respostas de referência, indicando o grau de equivalência de significado entre ambas.
- *Execution Rate*: proporção de consultas processadas com sucesso pelo *pipeline*, considerando as execuções concluídas sem falhas durante as etapas de

recuperação, processamento e geração da resposta.

- *Grounding Accuracy*: proporção de respostas cujas informações estão integralmente fundamentadas nas fontes institucionais recuperadas, indicando a capacidade do sistema de evitar informações não suportadas ou alucinações.
- *Avg Source IoU* (*Average Source Intersection over Union*): média da sobreposição entre o conjunto de documentos recuperados pelo sistema e o conjunto de documentos considerados corretos para cada consulta, refletindo a qualidade do mecanismo de recuperação de informações.

Para verificar o mecanismo de revisão de crenças, foram executadas duas baterias de testes. A primeira consistiu em 41 cenários sintéticos, elaborados para exercitar os estados **INALTERADA**, **ATUALIZADA** e **OBSOLETA**, incluindo casos de fronteira próximos aos limiares de decisão adotados pelo sistema. A segunda utilizou documentos institucionais reais submetidos a modificações controladas, simulando situações de atualização, manutenção e obsolescência do conhecimento. Essa etapa teve como objetivo verificar a consistência das decisões relacionadas à atualização da memória institucional.

Ademais, foram realizados testes voltados à verificação da infraestrutura de software. Nesse contexto, foram conduzidos testes de integração da *Application Programming Interface* (API), conjunto de interfaces que permite a troca de informações entre diferentes partes do sistema, verificando o funcionamento correto dos *endpoints*, a persistência dos dados e o gerenciamento de histórico e *feedback* dos usuários. Com o objetivo de garantir a reprodutibilidade dos experimentos e reduzir a dependência de serviços externos, foram empregadas técnicas de *mocking*, nas quais componentes reais são substituídos por versões simuladas durante a execução dos testes. Adicionalmente, foram realizados testes unitários para verificação isolada dos componentes e testes de contrato entre agentes, com o objetivo de assegurar a compatibilidade estrutural dos dados trocados ao longo do *pipeline*. Esses testes verificaram a presença, a consistência e a conformidade dos campos necessários à comunicação entre as diferentes etapas do fluxo.

3.1.2 Questionário e Testes com Usuários

Complementarmente à avaliação técnica, foi conduzida uma avaliação com usuários reais, com o objetivo de analisar a percepção da comunidade acadêmica em relação ao sistema desenvolvido. A coleta de dados ocorreu em duas etapas distintas. Na primeira, o *chatbot* foi disponibilizado a docentes do IFPI Campus Piripiri, permitindo a identificação de inconsistências e problemas de funcionamento antes de uma divulgação mais ampla. Com base nos relatos obtidos nessa fase, foram realizados ajustes relacionados principalmente à cobertura da base documental, à formatação

das fontes apresentadas nas respostas e a problemas de renderização da interface. Na segunda etapa, após a implementação dessas correções, o sistema foi disponibilizado aos discentes e demais membros da comunidade acadêmica, observando-se uma recepção mais positiva em relação à experiência de uso.

O instrumento de coleta adotado foi um questionário estruturado baseado no *Chatbot Usability Questionnaire* (CUQ), proposto por Holmes *et al.* (2019), com adaptações para o contexto institucional do IFPI. O CUQ original é composto por 16 questões balanceadas, sendo oito relacionadas a aspectos positivos e oito a aspectos negativos da usabilidade, avaliadas em escala Likert de cinco pontos. A pontuação final é calculada em uma escala de 0 a 100, permitindo a comparação objetiva da percepção de usabilidade entre diferentes sistemas conversacionais. Além das questões originais do instrumento, foram incluídas cinco questões adicionais destinadas a avaliar características específicas da solução proposta:

1. Correção das informações fornecidas em relação ao Manual do Servidor;
2. Consistência das respostas apresentadas pelo sistema;
3. Capacidade de combinar informações provenientes da documentação institucional e de fontes externas da internet;
4. Rastreabilidade das fontes utilizadas para fundamentar as respostas;
5. Percepção do usuário sobre a facilidade de acesso às informações institucionais proporcionada pelo *chatbot*.

O questionário incluiu ainda questões abertas e fechadas complementares relacionadas ao perfil dos participantes (tipo de usuário, curso ou setor de atuação e nível de familiaridade com tecnologia), ao objetivo da consulta realizada, ao sucesso na obtenção das informações desejadas, à identificação de eventuais erros durante o uso e à avaliação geral do sistema. Também foram coletados indicadores de intenção de reuso e recomendação da ferramenta.

O formulário foi disponibilizado em formato digital por meio do Google Forms, permanecendo aberto entre os dias 27 de maio e 12 de junho de 2026. Ao final do período de coleta, foram registradas 34 respostas válidas, sendo 10 provenientes de docentes e 24 de discentes ou outros membros da comunidade acadêmica.

A participação foi voluntária e o formulário apresentava, em sua seção inicial, os objetivos da pesquisa e a finalidade exclusivamente acadêmica dos dados coletados. Não foram coletados dados pessoais identificáveis, restringindo-se a coleta às questões do instrumento CUQ e a informações de perfil de caráter agregado, como vínculo institucional e nível de familiaridade com tecnologia. Os históricos de conversa registrados pela aplicação foram utilizados apenas para depuração técnica e não integraram o conjunto de dados analisado.

Os dados quantitativos foram analisados por meio de estatística descritiva, incluindo o cálculo de médias, medianas, desvios padrão e distribuições de frequência. A pontuação do CUQ foi calculada conforme o método oficial proposto pelos autores do instrumento. Complementarmente, as respostas abertas foram submetidas a uma análise qualitativa, buscando identificar padrões recorrentes relacionados a sugestões de melhoria, dificuldades encontradas e limitações percebidas pelos participantes durante a utilização do sistema.

4 SOLUÇÃO PROPOSTA

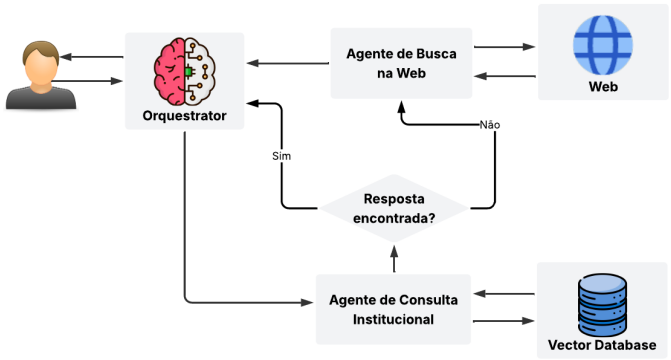
Esta seção apresenta a solução conversacional desenvolvida, descrevendo sua arquitetura, o fluxo de resposta às consultas dos usuários, o *pipeline* de atualização da base de conhecimento e os agentes responsáveis pelo processamento e pela revisão das informações institucionais. Também são apresentadas as tecnologias empregadas na implementação e as estratégias adotadas para ampliar a segurança e a confiabilidade do sistema.

4.1 ARQUITETURA E FLUXO DE RESPOSTA DO SISTEMA

A solução proposta foi desenvolvida com base no paradigma de Sistemas Multiagentes, adotando uma arquitetura composta por agentes especializados que atuam de forma coordenada na recuperação de informações e na geração de respostas aos usuários. A organização modular dos componentes possibilita a separação de responsabilidades específicas, favorecendo a manutenção, a escalabilidade e a evolução da solução.

O fluxo de processamento das consultas é apresentado na Figura 2. O processo é iniciado quando o usuário submete uma pergunta em linguagem natural ao sistema. Essa consulta é recebida pelo agente Orquestrador, responsável por coordenar a comunicação entre os demais agentes e controlar o fluxo de execução.

Figura 2 – Fluxograma de Respostas.



Fonte: Elaborado pelo autor.

Após o recebimento da solicitação, o Orquestrador encaminha a consulta diretamente ao Agente de Consulta Institucional, responsável por realizar a busca na base de conhecimento. Essa base é composta por documentos institucionais previamente processados e indexados em um banco vetorial, permitindo a recuperação de informações por similaridade semântica.

Concluída a etapa de busca, o sistema verifica se a base institucional efetivamente contém a informação solicitada, por meio de duas etapas complementares. Primeiro, os trechos recuperados do banco vetorial são filtrados por relevância semântica, de modo a selecionar os fragmentos mais próximos da pergunta. Em seguida, esses trechos são analisados pelo modelo de linguagem, que verifica se de fato respondem ao que foi perguntado. Quando o conteúdo recuperado sustenta uma resposta fundamentada, o resultado é retornado ao Orquestrador, que utiliza as informações recuperadas para compor a resposta final encaminhada ao usuário.

Caso a busca institucional não encontre informações suficientes ou a consulta esteja relacionada a conteúdos não presentes na base documental, o fluxo é redirecionado ao Agente de Busca na Web. Esse agente realiza consultas em fontes externas de informação, ampliando o escopo da recuperação de conhecimento para além dos documentos institucionais armazenados localmente. Após a execução da busca externa, os conteúdos relevantes encontrados são enviados ao Orquestrador, que os utiliza na elaboração da resposta final, caso nenhuma informação relevante seja encontrada ele retorna a falha ao obter a resposta. Dessa forma, a solução adota uma estratégia hierárquica de recuperação de informações, priorizando inicialmente a base institucional e recorrendo à Web apenas quando necessário.

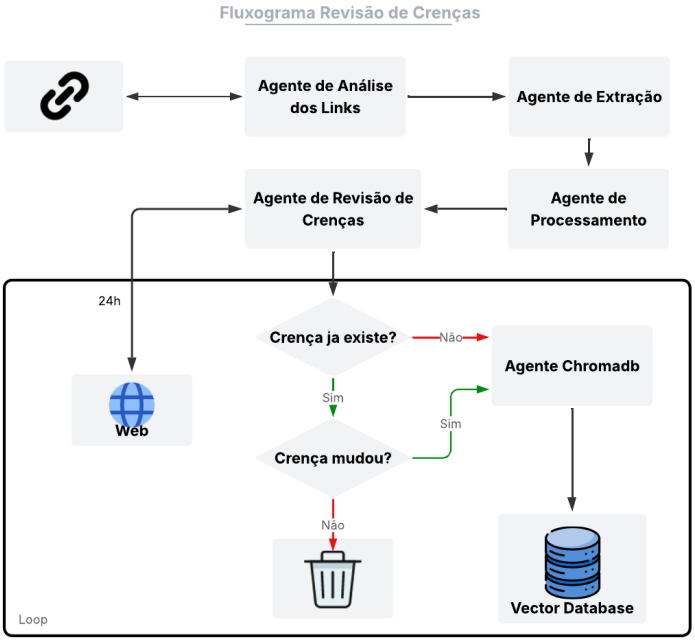
Essa abordagem contribui para preservar a confiabilidade das respostas relacionadas ao domínio institucional, ao mesmo tempo em que amplia a cobertura da solução para consultas que extrapolam o conteúdo presente na base de conhecimento local. Como resultado, obtém-se um mecanismo de resposta mais robusto e flexível, capaz de combinar conhecimento especializado da instituição com informações externas quando apropriado.

4.2 *PIPELINE* DE ATUALIZAÇÃO E REVISÃO DE CRENÇAS

Paralelamente ao fluxo de resposta, o sistema implementa um *pipeline* contínuo de atualização da base de conhecimento, responsável por garantir a consistência e a atualidade das informações institucionais. O processo é apresentado na Figura 3, que descreve o fluxo de revisão de crenças executado de forma recorrente em ciclos de 24 horas.

O *pipeline* é composto por cinco agentes especializados organizados de forma sequencial, cujas funcionalidades são detalhadas nas subseções seguintes: o **Agente de Análise de Links**, encarregado da validação e classificação das fontes institucio-

Figura 3 – Fluxograma de Revisão de crenças.



Fonte: Elaborado pelo autor.

nais; o **Agente de Extração**, responsável pela conversão dos documentos em texto estruturado; o **Agente de Processamento**, responsável pela segmentação semântica e geração dos *embeddings*; o **Agente de Revisão de Crenças**, que constitui o núcleo decisório do *pipeline*; e o **Agente ChromaDB**, responsável pela execução das decisões na camada de persistência vetorial.

O núcleo do *pipeline* corresponde ao Agente de Revisão de Crenças, responsável por avaliar a consistência das informações ao longo do tempo. Inspirado na teoria de revisão de crenças, esse agente compara, por meio da similaridade de cosseno, os vetores de *embedding* da versão mais recente indexada (*crença anterior*) com os vetores gerados a partir do documento recém-coletado (*crença nova*). A tomada de decisão segue uma sequência hierárquica de verificações:

1. caso a fonte do documento permaneça inacessível ou seja sinalizada como inválida mesmo após as tentativas alternativas de acesso descritas na Seção 4.2.1, a crença é classificada como *obsoleta* e marcada para sinalização ao administrador, sem remoção automática. Indisponibilidades contornáveis, como respostas de acesso restrito a consultas automatizadas, não acionam essa classificação, sendo tratadas pelas estratégias de recuperação do Agente de Análise de Links;

2. caso não exista uma crença anterior registrada (primeira indexação), não há base de comparação semântica; a crença é classificada como *atualizada*, com justificativa técnica registrada de indexação inicial, e seus *chunks* são encaminhados diretamente para indexação;

3. caso a média das similaridades de cosseno entre os *embeddings* antigos e novos seja inferior a 0,95, ou o volume de *chunks* apresente variação superior a 20% em relação à versão anterior, a crença é classificada como *atualizada*. A média é calculada tomando-se, para cada fragmento da versão anterior, a maior similaridade encontrada entre os fragmentos da versão recém-coletada, sem pressupor correspondência posicional entre eles. O grau da alteração é determinado com base nesse mesmo valor de similaridade média, sendo considerado *alto* quando inferior a 0,85 e *moderado* quando situado entre 0,85 e 0,95;

4. caso contrário, isto é, quando a similaridade for igual ou superior a 0,95 e o volume de *chunks* permanecer estável, a crença é classificada como *inalterada* e mantida sem modificações.

Nesse sentido, o ícone de descarte representado na Figura 3 refere-se ao descarte do conteúdo recém-coletado que não representa alteração semântica relevante, e não à exclusão da crença armazenada — que permanece inalterada na base de conhecimento institucional.

Ressalta-se que os limiares de 0,95, 0,85 e 20% não decorrem de valores de referência consolidados na literatura para esse contexto, tendo sido estabelecidos por calibração empírica sobre documentos institucionais do IFPI. Tal procedimento mostrou-se necessário porque documentos normativos apresentam estrutura formal e vocabulário recorrente, o que produz similaridades residuais elevadas mesmo entre versões distintas: limiares menos restritivos classificariam como inalteradas versões que apresentam alterações efetivas de conteúdo. Os valores adotados buscam, assim, equilibrar a detecção dessas alterações e a tolerância a mudanças não substantivas, como reformatações e correções ortográficas.

Além disso, os documentos utilizados na calibração dos limiares pertencem ao mesmo acervo institucional empregado na verificação descrita na Seção 5.2.2. Dessa forma, os resultados daquela etapa devem ser interpretados como verificação da correta implementação das regras de decisão sobre o domínio para o qual foram ajustadas, e não como estimativa de generalização do mecanismo a documentos não observados durante a calibração. A verificação controlada apresentada na Seção 5.2.1, por sua vez, não compartilha essa limitação, uma vez que utilizou vetores sintéticos construídos manualmente, com valores de similaridade previamente definidos.

Quando uma crença é classificada como *atualizada*, os novos *chunks* são encaminhados ao Agente ChromaDB, que substitui a versão anterior no banco de dados vetorial. Por outro lado, quando a crença é classificada como *inalterada*, nenhuma ação é executada, e o ciclo é reiniciado após o intervalo de 24 horas. O processo caracteriza-se, portanto, como contínuo e cíclico, permitindo a construção de uma base de conhecimento dinâmica, atualizada automaticamente sem a necessidade de

reindexação manual completa. Como consequência, reduz-se a ocorrência de inconsistências e aumenta-se a confiabilidade das respostas geradas pelo sistema.

4.2.1 Agente de Análise de Links

O **Agente de Análise de Links** constitui o ponto de entrada do *pipeline*, sendo responsável pela identificação e classificação das fontes documentais a serem processadas. Para isso, disponibiliza duas funcionalidades principais: a recuperação de todos os documentos oficiais cadastrados como ativos na base de dados e a análise individual de cada URL identificada. Quando a fonte não aponta diretamente para um arquivo nos formatos PDF, DOCX ou TXT, mas sim para uma página HTML, é realizado o carregamento e a varredura dos hiperlinks nela presentes, os quais são agregados como novas fontes passíveis de verificação.

A validação da acessibilidade das fontes é realizada por meio de tentativas de acesso ao conteúdo, em vez de verificações apenas dos metadados da página. Essa abordagem foi adotada porque alguns servidores institucionais bloqueiam determinados tipos de consulta, o que poderia levar à classificação incorreta de documentos disponíveis como inacessíveis. Além disso, quando um arquivo PDF retorna uma resposta de acesso restrito (**403**), a fonte não é descartada. Nesses casos, ela é classificada como **PDF_DIRETO** e encaminhada ao Agente de Extração, que realiza novas tentativas de obtenção do conteúdo utilizando métodos alternativos mais robustos. Somente quando essas estratégias alternativas também falham é que a fonte é considerada efetivamente inacessível, condição que subsidia a classificação de obsolescência realizada pelo Agente de Revisão de Crenças.

A classificação do conteúdo é realizada com base no tipo de mídia informado pelo servidor web, o qual indica o formato do recurso disponibilizado. Essa etapa permite distinguir diferentes categorias de fontes documentais, incluindo arquivos disponibilizados diretamente nos formatos PDF, DOCX e TXT, bem como páginas HTML que demandam processamento adicional.

Após validar a acessibilidade da fonte, o sistema identifica onde o conteúdo relevante está localizado. Em alguns casos, as informações institucionais encontram-se diretamente no texto da página HTML; em outros, a página funciona apenas como um repositório que referencia documentos externos, geralmente em formato PDF. Essa distinção é necessária para definir a estratégia de extração mais adequada.

Para realizar essa identificação, o sistema analisa a página em busca de links para arquivos PDF e verifica indícios de conteúdo institucional por meio do título e de trechos do texto, considerando termos como “resolução”, “portaria”, “instrução normativa” e “edital”. Com base nessa análise, as fontes são classificadas em uma das seguintes categorias:

- **HTML_COM_PDF**: páginas HTML que contêm links para documentos PDF, indi-

cando que o conteúdo relevante está armazenado em arquivos externos;

- **HTML_TEXTO**: páginas HTML cujo conteúdo relevante está presente diretamente no texto da página e possui volume informacional suficiente para processamento;
- **PDF_DIRETO**: endereços que apontam diretamente para arquivos PDF.

Essa classificação é utilizada para definir a estratégia de extração mais adequada e encaminhar a fonte para a etapa subsequente do *pipeline*.

4.2.2 Agente de Extração

O **Agente de Extração** é responsável por converter os documentos identificados pelo agente anterior em uma representação textual estruturada e adequada ao processamento computacional. Durante esse processo, são preservados elementos relevantes do documento, como títulos, parágrafos e tabelas, enquanto informações de formatação sem valor semântico são removidas. O conteúdo resultante é enriquecido com metadados, como idioma, tipo de documento e informações sobre o processo de extração.

A estratégia de extração é definida com base na classificação produzida pelo Agente de Análise de Links. Para documentos PDF, são empregados mecanismos que permitem recuperar texto, tabelas e a ordem lógica de leitura dos elementos presentes no documento. Além disso, uma heurística é utilizada para identificar documentos digitalizados (*scanned*), baseada na proporção de páginas que apresentam conteúdo textual extraível, indicando a necessidade potencial de reconhecimento óptico de caracteres.

Arquivos DOCX, TXT e páginas HTML são processados por métodos específicos para cada formato, visando recuperar o conteúdo textual relevante e eliminar elementos que não contribuem para a representação semântica do documento. No caso das tabelas, são geradas representações estruturadas e textuais para preservar as informações nelas contidas.

Independentemente do formato processado, o conteúdo extraído passa por uma etapa de normalização textual, que inclui a remoção de espaçamentos redundantes e a padronização das quebras de linha. Em seguida, é realizada uma verificação mínima de conteúdo útil, estabelecida em 100 caracteres, com o objetivo de descartar extrações vazias, incompletas ou provenientes de páginas de erro capturadas indevidamente. Documentos que não atendem a esse critério são reportados como falha e não seguem para as etapas subsequentes do *pipeline*.

4.2.3 Agente de Processamento

O **Agente de Processamento** transforma o texto estruturado produzido pelo agente anterior em unidades semânticas menores (*chunks*) e gera suas representações vetoriais (*embeddings*), etapa essencial para a recuperação de informações por similaridade semântica no módulo de RAG. A segmentação é realizada com o RecursiveCharacterTextSplitter, da biblioteca LangChain, configurado para priorizar separadores típicos de documentos normativos, como títulos, capítulos, seções e artigos, preservando a estrutura lógica e reduzindo a fragmentação de informações relacionadas.

Os *chunks* são gerados com aproximadamente 800 caracteres e sobreposição de 150 caracteres entre *chunks*, buscando preservar o contexto entre trechos adjacentes. Além disso, quando necessário, o título do documento é adicionado como cabeçalho ao fragmento, garantindo que o contexto documental seja incorporado às representações vetoriais e às etapas posteriores de recuperação. Após a segmentação, são gerados e normalizados os *embeddings* de cada fragmento, tornando as medidas de similaridade dependentes principalmente do conteúdo semântico. Por fim, cada *chunk* é enriquecido com metadados de rastreabilidade, como origem e posição no documento, resultando em fragmentos vetorizados e contextualizados que servem como unidade básica para a indexação e atualização da base de conhecimento.

4.2.4 Agente de Revisão de Crenças

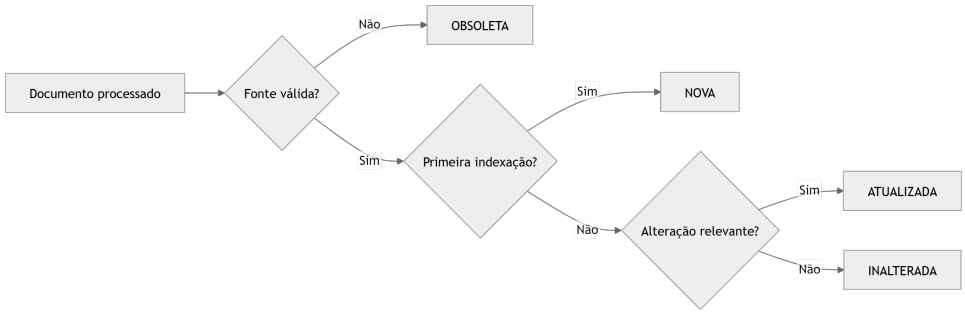
O **Agente de Revisão de Crenças** constitui o núcleo decisório do *pipeline*, sendo responsável por avaliar se o conhecimento previamente armazenado deve ser mantido, atualizado ou sinalizado como potencialmente obsoleto a partir das evidências fornecidas pelos agentes anteriores. Inspirado nos princípios da teoria de Revisão de Crenças AGM, esse componente atua como um mecanismo de manutenção contínua da base de conhecimento institucional.

Como entrada, o agente recebe os *chunks* produzidos pelo Agente de Processamento e, quando existentes, as representações vetoriais da versão previamente indexada do documento. A partir desses dados, realiza a comparação entre as representações semânticas por meio da similaridade de cosseno, além de considerar informações estruturais do documento, como variações na quantidade de *chunks*. Os critérios e limiares utilizados nessa análise foram apresentados na Seção 4.2.

Ao término da avaliação, o agente produz uma decisão indicando o estado da crença (**INALTERADA**, **ATUALIZADA** ou **OBSOLETA**), acompanhada de metadados e de uma justificativa técnica gerada automaticamente. Essa justificativa registra os indicadores considerados durante a análise e os motivos que fundamentaram a decisão, fornecendo mecanismos de explicabilidade, auditoria e rastreabilidade. A Figura 4 apresenta, de forma resumida e visual, o fluxo de decisão utilizado pelo Agente de

Revisão de Crenças para classificar o estado do conhecimento institucional.

Figura 4 – Fluxo simplificado de decisão do Agente de Revisão de Crenças.



Fonte: Elaborado pelo autor.

A decisão produzida é então encaminhada ao Agente ChromaDB, responsável por executar as operações correspondentes na base vetorial. Quando a crença é classificada como **ATUALIZADA**, a versão anterior é substituída pela nova representação; quando classificada como **INALTERADA**, nenhuma modificação é realizada; e, nos casos de **OBSOLETA**, o documento permanece preservado até a validação por um administrador, conforme o fluxo de tratamento descrito na Seção 4.2.5.

4.2.5 Agente ChromaDB

O **Agente ChromaDB** é responsável por materializar, na camada de persistência, as decisões produzidas pelo Agente de Revisão de Crenças, mantendo a memória institucional consistente por meio das operações de preservação, atualização ou sinalização dos conhecimentos armazenados no banco vetorial.

As ações executadas correspondem diretamente ao estado atribuído pelo Agente de Revisão de Crenças. Quando a crença é classificada como *inalterada*, nenhuma modificação é realizada (**MANTER**). Quando classificada como *atualizada*, os vetores da versão anterior são substituídos pelos novos fragmentos processados (**ATUALIZAR**), garantindo que a base reflita a versão mais recente do documento. Embora exista suporte para a ação **REMOVER**, ela é utilizada apenas em testes, pois a permanência do documento na base de conhecimento é condição necessária para que o próprio mecanismo de revisão possa, em ciclos futuros, comparar a versão armazenada com novas coletas e detectar alterações semânticas, sinalizando a mudança ao usuário e ao administrador. Caso a crença fosse removida automaticamente, o sistema perderia a referência necessária para identificar que houve mudança, comprometendo a lógica de revisão contínua proposta neste trabalho.

Quando um documento é classificado como *obsoleta*, o agente executa a ação **SINALIZAR_OBSOLESCENCIA**, iniciando um fluxo em duas etapas. Inicialmente, o documento é apenas sinalizado ao administrador, permanecendo ativo na base. A confirmação definitiva depende de decisão humana, que marca o documento como

obsoleto e o remove das rotinas de monitoramento, evitando descontinuações indevidas causadas por indisponibilidades temporárias da fonte.

Como limitação atual, documentos confirmados como obsoletos permanecem armazenados no ChromaDB e ainda podem ser recuperados pelo mecanismo de RAG, pois não há filtragem por metadados durante a busca semântica. A implementação desse filtro foi identificada como um aprimoramento para trabalhos futuros.

Antes da persistência, os metadados dos fragmentos são adequados ao esquema do banco vetorial, garantindo compatibilidade e integridade das informações. Dessa forma, o agente assegura que as decisões da revisão de crenças sejam refletidas de maneira consistente na memória institucional utilizada pelo mecanismo de recuperação semântica.

4.3 TECNOLOGIAS UTILIZADAS

A implementação do sistema foi realizada com o emprego de modelos de linguagem baseados na arquitetura *Transformer*, amplamente utilizada em tarefas de PLN. Como modelo de linguagem principal, foi utilizado o modelo Gemini, da Google, por meio da versão gemini-3.1-flash-lite, responsável pela geração das respostas do sistema. A escolha dessa variante foi motivada pelo equilíbrio entre desempenho, custo operacional e baixa latência, características importantes para viabilizar a utilização do sistema em um ambiente institucional. Além disso, como a arquitetura proposta adota a abordagem RAG, a recuperação das informações relevantes é realizada a partir da base documental antes da geração da resposta, reduzindo a dependência do conhecimento paramétrico do modelo e tornando adequada a utilização de uma variante mais leve para o contexto deste trabalho. O *backend* da aplicação foi desenvolvido em Python, utilizando o *framework* Flask para a camada web e para a exposição das rotas de interação. A persistência de dados relacionais, incluindo informações de usuários, histórico de conversas e metadados de documentos, foi realizada por meio do banco de dados PostgreSQL.

A orquestração dos agentes foi implementada com o auxílio dos *frameworks* LangChain e LangGraph. O LangChain foi empregado para integrar o modelo de linguagem aos mecanismos de recuperação de informação, bancos de dados vetoriais e demais componentes da aplicação. Já o LangGraph foi escolhido por oferecer controle explícito do fluxo de execução por meio de grafos de estados, permitindo implementar ciclos, ramificações condicionais e mecanismos de decisão adequados à arquitetura multiagente proposta. Além disso, sua integração nativa com o ecossistema do LangChain e sua maturidade no ambiente Python favoreceram sua adoção em detrimento de alternativas como AutoGen e CrewAI, que oferecem menor controle sobre a lógica de orquestração dos agentes. Adicionalmente, a aplicação foi conteinerizada com Docker, viabilizando a reprodutibilidade do ambiente de execução e a separação de

responsabilidades entre os diferentes serviços que compõem a arquitetura.

Para o armazenamento das representações vetoriais dos documentos institucionais foi adotado o ChromaDB (Xie *et al.*, 2023). A escolha foi motivada por sua integração nativa com o ecossistema LangChain, pela facilidade de implantação e pela adequação a aplicações baseadas em bases documentais de pequeno e médio porte, como a utilizada neste trabalho. Além disso, por ser uma solução gratuita e de código aberto, o ChromaDB reduz o custo operacional e elimina a necessidade de infraestrutura especializada durante as fases de desenvolvimento e validação do sistema. Embora existam alternativas consolidadas, como Pinecone, Weaviate e pgvector, a escolha deste trabalho priorizou simplicidade de integração, rapidez de prototipagem e facilidade de manutenção, em vez de requisitos de escalabilidade para ambientes de produção em larga escala. Dessa forma, não se buscou demonstrar superioridade de desempenho em relação a outros bancos vetoriais, mas selecionar uma tecnologia compatível com os objetivos, o porte da base de conhecimento e os recursos disponíveis para o desenvolvimento do protótipo.

Os *embeddings* são gerados por meio do modelo multilíngue paraphrase-multilingual-MiniLM-L12-v2, pertencente à família Sentence-Transformers. O modelo foi selecionado por oferecer suporte nativo a múltiplos idiomas, incluindo o português brasileiro, e por apresentar um equilíbrio entre qualidade semântica e custo computacional. Essa característica possibilita sua execução local sem dependência de serviços externos ou infraestrutura dedicada de aceleração gráfica. Dessa forma, o processo de vetorização pode ser realizado de maneira eficiente, preservando a autonomia da solução e reduzindo seus custos operacionais.

4.4 SEGURANÇA E PROTEÇÃO CONTRA *PROMPT INJECTION*

Sistemas baseados em modelos de linguagem estão sujeitos a ataques de *prompt injection*, nos quais instruções maliciosas inseridas em documentos ou consultas podem influenciar o comportamento do modelo, levando à geração de respostas inadequadas, incorretas ou incompatíveis com as regras estabelecidas pela aplicação. Esse risco torna-se particularmente relevante em arquiteturas baseadas em RAG, nas quais conteúdos recuperados da base de conhecimento são incorporados ao contexto enviado ao modelo de linguagem.

Com o objetivo de mitigar esse problema, foi implementado um mecanismo de defesa composto por duas camadas complementares. A primeira consiste na definição explícita de instruções de segurança no *prompt* do sistema, orientando o modelo a tratar os documentos recuperados exclusivamente como fontes de informação e a ignorar quaisquer comandos, instruções ou tentativas de alteração de comportamento presentes em seu conteúdo. Dessa forma, busca-se reduzir a influência de trechos que possam ser interpretados como instruções direcionadas ao modelo.

A segunda camada foi implementada diretamente no código da aplicação por meio de um processo de sanitização dos documentos recuperados antes de sua utilização no contexto de geração das respostas. Essa etapa utiliza expressões regulares para identificar e remover padrões potencialmente associados a ataques de *prompt injection*, tais como comandos explícitos direcionados ao modelo, solicitações para ignorar instruções anteriores e outras construções linguísticas frequentemente empregadas para manipular o comportamento de sistemas conversacionais.

A utilização conjunta dessas estratégias caracteriza uma abordagem de defesa em profundidade (*defense in depth*), princípio amplamente adotado na área de segurança da informação, segundo o qual a proteção de um sistema não deve depender de um único mecanismo de segurança, mas de múltiplas camadas complementares (Holmberg, 2017). Assim, mesmo que uma instrução maliciosa não seja detectada durante a sanitização, ela poderá ser bloqueada pelas restrições definidas no *prompt* do sistema. Da mesma forma, conteúdos perigosos que escapem a essas restrições podem ser mitigados pela filtragem prévia dos documentos. A combinação dessas camadas aumenta a robustez da aplicação e reduz a probabilidade de sucesso de ataques de *prompt injection*.

4.5 DISPONIBILIDADE DO ARTEFATO

Com o propósito de assegurar a verificabilidade dos procedimentos descritos nesta seção, o código-fonte da solução, os arquivos de conteinerização, a suíte de testes automatizados e o conjunto de questões utilizado na avaliação técnica encontram-se disponíveis em repositório público[3]. O repositório inclui as instruções de execução do ambiente via Docker, os *scripts* de avaliação e os arquivos de teste referentes às verificações apresentadas na Seção 5. Ressalta-se que as chaves de acesso aos serviços externos não são distribuídas, devendo ser configuradas pelo usuário, e que a reprodução integral dos resultados de geração de respostas é limitada pela natureza não determinística do modelo de linguagem empregado, conforme discutido na Seção 5.5.

5 RESULTADOS E DISCUSSÃO

Esta seção apresenta os resultados obtidos a partir da avaliação técnica do sistema e da percepção dos usuários, com o objetivo de analisar sua eficácia na interpretação de normativas institucionais e na manutenção da consistência do conhecimento.

[3]Disponível em: https://github.com/larissaNa/chatbot-ifpi. Acesso em: 22 jul. 2026.

5.1 RESULTADOS DA AVALIAÇÃO TÉCNICA

A avaliação técnica foi conduzida com base em um conjunto de 51 questões representativas do domínio institucional, contemplando consultas sobre regulamentos acadêmicos, informações administrativas, calendários, horários de cursos e normas funcionais. Cada item do conjunto de avaliação foi composto por uma pergunta, uma resposta de referência e uma ou mais fontes esperadas.

A avaliação foi realizada com base em cinco métricas complementares, voltadas à análise da corretude semântica das respostas, da robustez operacional do sistema e da legitimidade das fontes recuperadas: *Semantic Accuracy*, *Avg Semantic Similarity*, *Execution Rate*, *Grounding Accuracy* e *Avg Source IoU*. Os resultados globais obtidos estão apresentados na Tabela 1.

Tabela 1 – Resultados da avaliação técnica.

Métrica	Valor
Semantic Accuracy	0,63
Avg Semantic Similarity	0,78
Execution Rate	1,00
Grounding Accuracy	0,78
Avg Source IoU	0,54

Fonte: Elaborado pelo autor.

A taxa de execução de 1,00 indica que todas as 51 perguntas foram processadas com sucesso, sem falhas de *pipeline*, *timeouts* ou erros de integração com serviços externos. Esse resultado evidencia a estabilidade operacional da arquitetura proposta, garantindo que a avaliação de conteúdo não foi comprometida por falhas técnicas durante o processamento.

A *Grounding Accuracy* foi apurada de forma automática, comparando programaticamente as fontes anexadas às respostas geradas com as fontes de referência definidas para cada questão. Considerou-se uma resposta fundamentada quando ao menos uma das fontes correspondia, total ou parcialmente, a uma fonte esperada ou pertencia ao domínio institucional oficial do IFPI, indicando a recuperação de um documento legítimo. Esse procedimento, baseado exclusivamente em comparação de *strings*, garante reprodutibilidade e elimina subjetividade na avaliação da métrica.

A *Grounding Accuracy* de 0,78 mostra que cerca de quatro em cada cinco respostas geradas puderam ser associadas a fontes institucionais legítimas. Esse resultado é particularmente relevante no contexto da arquitetura proposta, que combina recuperação local em base vetorial com escalonamento para busca web quando a evidência disponível no repositório indexado não é suficiente. Em outras palavras, o sistema demonstrou capacidade consistente de fundamentar suas respostas em documentos oficiais ou em páginas do domínio institucional, mesmo quando a informação não foi obtida diretamente da base vetorial.

Por outro lado, a *Avg Source IoU* de 0,54 revela que a correspondência exata entre as fontes recuperadas e as fontes esperadas ainda é moderada. Esse comportamento decorre do caráter estrito da métrica, que considera a igualdade entre os títulos das fontes retornadas e os títulos presentes no gabarito, penalizando situações em que o sistema recupera um documento oficial alternativo, semanticamente equivalente, ou um PDF institucional obtido via busca web cujo título não coincide exatamente com o nome do documento indexado localmente. Assim, a diferença entre *Grounding Accuracy* e *Avg Source IoU* sugere que o sistema frequentemente ancora a resposta em conteúdo legítimo, mas nem sempre no mesmo artefato documental explicitamente previsto no gabarito.

No que se refere à qualidade semântica das respostas, a *Semantic Accuracy* de 0,63 indica que aproximadamente dois terços das respostas (32 de 51) atingiram o limiar de similaridade estabelecido para classificação como corretas. Esse valor deve ser interpretado em conjunto com a *Avg Semantic Similarity* de 0,78, superior ao limiar de aceitação adotado (0,75). Essa combinação sugere que parte expressiva das respostas classificadas como incorretas não corresponde a alucinações completas ou desvios graves de conteúdo, mas sim a respostas semanticamente próximas do gabarito, penalizadas por omissões pontuais, síntese excessiva ou diferenças de formulação textual.

A análise individual dos itens reforça essa interpretação: das 19 respostas classificadas como incorretas, 12 apresentaram similaridade entre 0,70 e 0,75, situando-se imediatamente abaixo do limiar de corte, e diversas delas foram devidamente ancoradas nas fontes esperadas, com correspondência integral de títulos. Nesses casos, o conteúdo essencial da resposta estava presente, mas diferenças de granularidade ou de redação em relação à resposta de referência reduziram o valor de similaridade dos *embeddings*. Observou-se ainda que os erros se concentram em questões sobre documentos normativos extensos, nos quais o trecho que contém a resposta exata nem sempre figura entre os fragmentos mais bem ranqueados pela busca vetorial, levando o sistema a escalonar para mecanismos de contingência que produzem respostas mais genéricas.

De forma geral, os resultados demonstram que o sistema é robusto e capaz de fundamentar suas respostas em fontes institucionais confiáveis. As principais limitações observadas estão relacionadas à recuperação de trechos específicos em documentos longos e à correspondência exata das fontes em cenários com documentos semelhantes. Apesar desses desafios, os resultados indicam que a arquitetura proposta é adequada ao domínio institucional, embora ainda possa ser aprimorada para aumentar a consistência das respostas e das referências utilizadas.

5.2 VERIFICAÇÃO DO MECANISMO DE REVISÃO DE CRENÇAS

5.2.1 Verificação Controlada da Lógica de Revisão

A primeira etapa de verificação teve como objetivo verificar o comportamento lógico do mecanismo de revisão de crenças em cenários controlados. Para isso, foi desenvolvida uma bateria de testes automatizados composta por 41 casos sintéticos, projetados para exercitar os diferentes estados previstos pelo modelo de decisão: manutenção do conhecimento (**INALTERADA**), atualização (**ATUALIZADA**) e obsolescência (**OBSOLETA**). Os cenários avaliados contemplaram diferentes situações previstas pelo mecanismo de revisão de crenças, incluindo:

- documentos sem alterações;
- pequenas modificações de conteúdo;
- alterações moderadas;
- alterações significativas;
- mudanças estruturais no volume de *chunks*;
- indisponibilidade das fontes documentais;
- casos de fronteira próximos aos limiares de decisão adotados pelo agente.

Para garantir controle preciso sobre os níveis de similaridade avaliados, foram utilizados vetores sintéticos com valores previamente definidos, permitindo verificar o comportamento do mecanismo em condições específicas de teste.

Os resultados obtidos são apresentados na Tabela 2. Em todos os cenários avaliados, o agente produziu a classificação esperada, totalizando 41 acertos em 41 casos executados. Observa-se que o mecanismo apresentou comportamento consistente tanto em situações convencionais quanto em condições próximas aos limiares de decisão, indicando estabilidade na aplicação das regras de revisão implementadas.

Os testes de fronteira merecem destaque por avaliarem situações próximas aos limiares utilizados pelo mecanismo para diferenciar documentos atualizados de documentos inalterados. Foram considerados cenários imediatamente acima e abaixo dos limites de similaridade e variação estrutural definidos pelo modelo, sendo observada classificação correta em todos os casos avaliados. Embora essa bateria não tenha sido concebida para estimar desempenho estatístico em larga escala, os resultados demonstram que a lógica de decisão implementada opera de forma consistente dentro dos critérios estabelecidos pelo sistema.

Por fim, vale ressaltar que a acurácia de 100% reflete a natureza determinística da lógica de decisão avaliada: como os limiares de classificação são definidos explicitamente no agente, o resultado esperado para cada cenário pode ser previsto a priori,

Tabela 2 – Resultados da verificação controlada do mecanismo de revisão de crenças.

Categoria de teste	**Casos**	**Acertos**	**Acurácia**
Documentos inalterados	5	5	100%
Pequenas alterações	5	5	100%
Alterações moderadas	5	5	100%
Alterações significativas	5	5	100%
Mudanças estruturais	6	6	100%
Fontes inacessíveis	5	5	100%
Casos de fronteira	10	10	100%
Total	**41**	**41**	**100%**

Fonte: Elaborado pelo autor.

e a verificação serve para confirmar a correta implementação dessas regras, e não para estimar uma taxa de erro probabilística do sistema.

5.2.2 Verificação com Documentos Institucionais Reais

Complementarmente à verificação controlada, foi conduzida uma avaliação utilizando documentos reais presentes na base de conhecimento do sistema. Diferentemente da bateria sintética, essa etapa exerceu o *pipeline* completo de processamento, incluindo segmentação em *chunks*, geração de *embeddings* pelo modelo Sentence-Transformers e execução do mecanismo de revisão de crenças.

Foram selecionados documentos institucionais efetivamente utilizados pelo *chatbot*, incluindo horários acadêmicos, escalas docentes e comunicados relacionados aos sábados letivos. Sobre esses documentos foram realizadas modificações controladas que simulam situações recorrentes no contexto institucional, como alteração de datas, substituição de disciplinas ou docentes, inclusão e remoção de conteúdo e indisponibilidade da fonte original.

Ao todo, foram executados 24 cenários distribuídos entre documentos inalterados, alterações de conteúdo, mudanças estruturais e situações de obsolescência. Os resultados são apresentados na Tabela 3. Em todos os casos, a classificação produzida pelo agente correspondeu ao resultado esperado, totalizando 24 acertos em 24 cenários avaliados.

Tabela 3 – Resultados da verificação com documentos institucionais reais.

Categoria de teste	**Casos**	**Acertos**	**Acurácia**
Documentos inalterados	6	6	100%
Alterações de conteúdo	6	6	100%
Mudanças estruturais	6	6	100%
Documentos obsoletos	6	6	100%
Total	**24**	**24**	**100%**

Fonte: Elaborado pelo autor.

Além da classificação correta dos estados de revisão, os testes permitiram verificar o funcionamento integrado do fluxo de atualização do conhecimento. Alterações de conteúdo foram corretamente identificadas como atualizações, documentos inalterados mantiveram seu estado original e fontes marcadas como indisponíveis foram classificadas como obsoletas. Dessa forma, os resultados fornecem evidências de que o mecanismo proposto é capaz de acompanhar alterações em documentos institucionais reais e refletir essas mudanças na base de conhecimento utilizada pelo *chatbot*.

Considerando conjuntamente as duas baterias de testes, foram avaliados 65 cenários distintos, abrangendo tanto condições controladas quanto situações próximas ao ambiente de produção. Os resultados obtidos reforçam a viabilidade da abordagem proposta para apoiar a manutenção contínua da memória institucional do sistema, permitindo preservar informações válidas, incorporar alterações relevantes e identificar potenciais situações de obsolescência de maneira consistente e rastreável.

5.3 TESTES DE INTEGRAÇÃO E CONSISTÊNCIA

Além da verificação específica do mecanismo de revisão de crenças, foi realizada uma bateria complementar de testes automatizados com o objetivo de verificar a consistência estrutural do sistema e a integração entre seus componentes. Esses testes abrangeram diferentes níveis de verificação, incluindo testes unitários, testes de integração, testes de contrato e testes funcionais das interfaces de comunicação da aplicação.

Os testes unitários foram utilizados para validar individualmente componentes como os agentes de análise de links, extração de conteúdo, processamento de documentos, revisão de crenças e persistência vetorial. Os testes de integração avaliaram o funcionamento conjunto desses componentes ao longo do *pipeline*, verificando a troca de informações entre agentes e a correta execução dos fluxos de processamento. Os testes de contrato tiveram como objetivo assegurar a compatibilidade estrutural dos dados produzidos e consumidos pelos agentes, enquanto os testes funcionais verificaram as rotas HTTP da aplicação, incluindo operações relacionadas ao envio de mensagens, histórico de conversas e gerenciamento de *feedbacks*.

Ao todo, foram executados 27 cenários de teste distribuídos entre as diferentes categorias de verificação. Conforme apresentado na Tabela 4, todos os cenários foram concluídos com sucesso, sem falhas observadas durante a execução da suíte.

A cobertura dos testes contemplou desde funcionalidades isoladas até fluxos completos da aplicação. Foram avaliados o processo de classificação de documentos pelo Agente de Análise de Links, os mecanismos de extração de conteúdo para diferentes formatos de arquivo, a geração de *chunks* e *embeddings*, o mecanismo de revisão de crenças, as operações de persistência no banco vetorial, o fluxo de geren-

Tabela 4 – Resultados dos testes de integração e consistência do sistema.

Categoria	**Cenários**	**Acertos**	**Taxa de sucesso**
Testes unitários	10	10	100%
Testes de integração	9	9	100%
Testes de contrato	3	3	100%
APIs e *endpoints* HTTP	5	5	100%
Total	**27**	**27**	**100%**

Fonte: Elaborado pelo autor.

ciamento de obsolescência de documentos e as interfaces HTTP disponibilizadas pelo sistema.

No total, a suíte compreendeu 27 cenários de teste automatizados distribuídos em 11 arquivos de teste, contendo 104 asserções explícitas. Nenhuma falha foi identificada durante a execução. Esses resultados fornecem evidências de que os componentes da arquitetura multiagente interagem de forma consistente e que os mecanismos de comunicação, persistência e atualização do conhecimento operam conforme esperado.

5.4 AVALIAÇÃO COM USUÁRIOS

A avaliação contou com a participação de 34 usuários. A Tabela 5 apresenta o perfil da amostra quanto ao vínculo institucional e ao nível de familiaridade com tecnologia. Os participantes utilizaram o sistema principalmente para consultas sobre calendário acadêmico, regulamentos institucionais, horários de aula e testes gerais de funcionamento.

Tabela 5 – Perfil dos participantes da avaliação.

Característica	n	%
Vínculo institucional		
Docentes	10	29,4
Discentes e outros usuários	24	70,6
Experiência tecnológica		
Básico	6	17,6
Intermediário	14	41,2
Avançado	14	41,2

Fonte: Elaborado pelo autor.

5.4.1 Resultados Quantitativos

A Tabela 6 apresenta os resultados consolidados da avaliação. Os discentes atribuíram avaliações superiores tanto na pontuação do CUQ quanto à nota geral. A nota geral média foi de 8,1 (desvio padrão = 1,4), sem registros de avaliações inferiores a 5. Além disso, a proximidade entre média e mediana do CUQ sugere ausência

de forte assimetria na distribuição das avaliações, enquanto o desvio padrão indica variabilidade moderada entre os participantes.

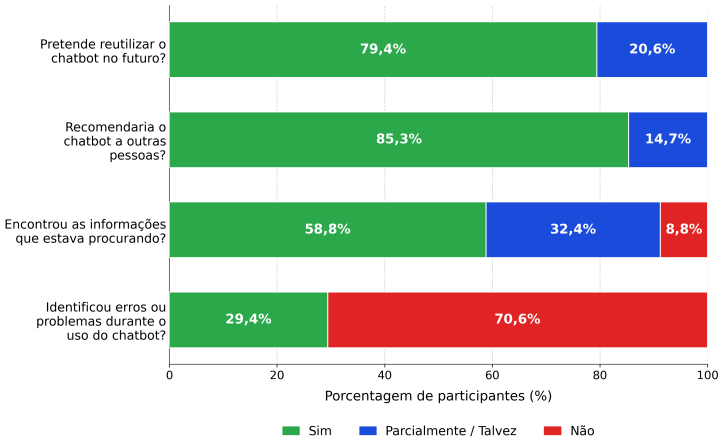
Tabela 6 – Resultados consolidados da avaliação com usuários.

Indicador	**Geral (n=34)**	**Docentes (n=10)**	**Discentes e outros (n=24)**
Pontuação CUQ (0–100)	**70,8**	62,2	74,4
Mediana CUQ (0–100)	71,9	59,4	78,9
Desvio padrão CUQ (0–100)	13,4	15,2	11,8
Nota geral (0–10)	**8,1**	7,5	**8,4**

Fonte: Elaborado pelo autor.

Indicadores gerais de aceitação Os indicadores gerais de aceitação, apresentados na Figura 5, evidenciam uma percepção positiva dos participantes em relação ao *chatbot*. Destacam-se os elevados índices de intenção de recomendação (85,3%) e de reutilização do sistema (79,4%). Além disso, a maioria dos usuários afirmou ter encontrado total ou parcialmente as informações desejadas, enquanto cerca de três quartos dos participantes afirmaram não ter identificado erros durante a utilização do sistema, reforçando a percepção de utilidade e de consistência das respostas fornecidas.

Figura 5 – Índices de interação e recomendação.

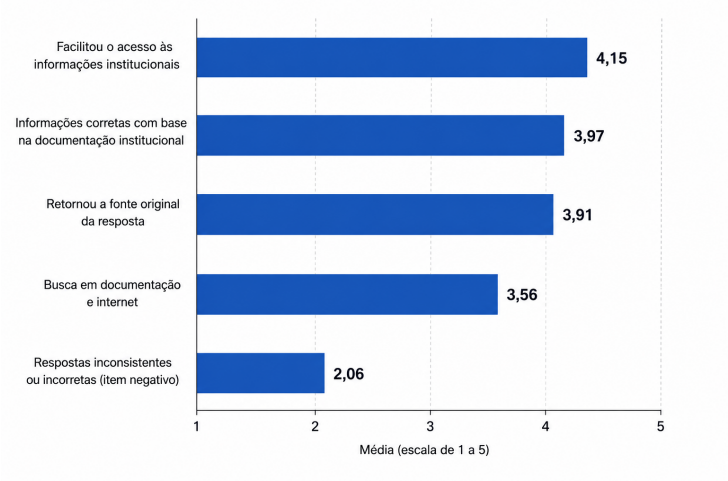


Fonte: Elaborado pelo autor.

Questões adicionais sobre características do sistema A Figura 6 apresenta as médias obtidas nas questões adicionais aplicadas aos participantes. As maiores médias foram registradas para a facilidade de acesso às informações institucionais (4,15), as respostas corretas (3,97) e a disponibilização das fontes utilizadas (3,91). Em contraste, a busca combinada entre documentos institucionais e fontes externas apresentou a menor média entre os itens positivos (3,56), indicando oportunidade de aprimo-

ramento. Já o item formulado negativamente sobre respostas inconsistentes obteve média de 2,06, reforçando a percepção de consistência das respostas geradas.

Figura 6 – Principais indicadores de aceitação e utilidade percebida do *chatbot*.



Fonte: Elaborado pelo autor.

Análise por dimensão do CUQ A análise das médias por questão do instrumento CUQ permitiu identificar as dimensões de maior e menor desempenho percebido pelos participantes. As maiores médias foram obtidas para a facilidade de uso do sistema (Q7: 4,21), a receptividade inicial durante a interação (Q3: 4,18) e a percepção de simplicidade e intuitividade (Q15: 4,18). A naturalidade da interação também recebeu avaliação elevada (Q1: 4,06), indicando que os usuários consideraram a experiência de uso agradável e acessível.

Entre os itens formulados negativamente, as menores médias foram observadas nas questões relacionadas à complexidade de uso (Q16: 1,79), à possibilidade de confusão durante a interação (Q8: 1,79) e à irrelevância das respostas fornecidas (Q12: 2,06). Como esses itens possuem escala invertida, tais resultados reforçam a percepção positiva dos participantes quanto à usabilidade e à qualidade das respostas geradas.

Por outro lado, algumas dimensões receberam avaliações relativamente inferiores quando comparadas aos demais aspectos analisados. Destacam-se a capacidade de lidar com erros ou perguntas mal formuladas (Q13: 3,50), a compreensão das perguntas dos usuários (Q9: 3,56) e a clareza na explicação do objetivo e funcionamento do sistema (Q5: 3,71). Embora essas médias permaneçam acima do ponto médio da escala, elas indicam oportunidades de aprimoramento relacionadas à interpretação de consultas mais complexas, ao tratamento de ambiguidades e à comunicação das funcionalidades do sistema.

Em conjunto, os resultados quantitativos evidenciam que o *chatbot* alcançou níveis satisfatórios de usabilidade e aceitação pelos participantes. A pontuação obtida

no CUQ, associada à elevada nota geral e aos altos índices de intenção de reutilização e recomendação, indica que a solução atende adequadamente às necessidades da maioria dos usuários. As oportunidades de melhoria concentram-se principalmente na compreensão de consultas ambíguas, no tratamento de perguntas mal formuladas e na comunicação mais clara das funcionalidades disponíveis, aspectos que poderão orientar a evolução do sistema em trabalhos futuros.

5.4.2 Análise Qualitativa das Respostas Abertas

A análise das respostas abertas permitiu identificar padrões recorrentes de percepção entre os participantes, organizados em dois eixos principais: pontos positivos e limitações observadas.

Entre os aspectos mais elogiados destacaram-se a facilidade de navegação e a intuitividade da interface, mencionadas por participantes com diferentes níveis de familiaridade tecnológica. A aderência do *chatbot* ao escopo institucional do IFPI também foi percebida como um diferencial. Nesse contexto, um participante relatou ter tentado induzir o sistema a responder questões fora do domínio institucional sem sucesso, afirmando que o *chatbot* “não apresentou alucinações”. Além disso, a rastreabilidade das fontes proporcionada pelo mecanismo de RAG foi amplamente valorizada pela possibilidade de verificar a origem das informações fornecidas.

A principal limitação identificada refere-se à cobertura da base documental, evidenciada por dificuldades em responder adequadamente a algumas perguntas relacionadas aos documentos indexados. Esse comportamento foi observado com maior frequência nas avaliações realizadas por docentes antes dos ajustes na base de conhecimento. Também foram mencionadas respostas excessivamente longas, motivando sugestões para torná-las mais objetivas. Entre as melhorias propostas destacam-se a integração com sistemas institucionais, como o SUAP, e a ampliação da base de conhecimento para incluir informações específicas de campi e turmas, como horários e cronogramas acadêmicos.

De forma geral, os resultados qualitativos corroboram os achados quantitativos, indicando boa usabilidade e confiabilidade percebida do sistema. Entretanto, evidenciam-se limitações relacionadas à cobertura da base de conhecimento e à necessidade de aprimoramentos na interface e na concisão das respostas, de modo a atender de forma mais adequada às expectativas da comunidade acadêmica.

5.5 DISCUSSÃO

Os resultados da avaliação técnica e da avaliação com usuários indicam que a integração entre geração de linguagem, recuperação de informação e revisão de crenças contribui para um sistema mais confiável e adequado ao contexto institucional. Na avaliação técnica, a arquitetura apresentou estabilidade operacional, processando to-

das as consultas e produzindo respostas semanticamente próximas aos gabaritos. A combinação entre *Semantic Accuracy* de 0,63 e *Avg Semantic Similarity* de 0,78 sugere que parte das respostas classificadas como incorretas continha conteúdo próximo ao esperado, sendo penalizada principalmente por diferenças de redação ou omissões.

A abordagem baseada em *Retrieval-Augmented Generation* mostrou boa capacidade de fundamentar as respostas em fontes institucionais, refletida pela *Grounding Accuracy* de 0,78. Por outro lado, a *Avg Source IoU* de 0,54 evidencia limitações na recuperação exata das fontes, especialmente em documentos extensos, nos quais o trecho contendo a informação desejada nem sempre é recuperado entre os primeiros resultados.

A arquitetura multiagente favoreceu a modularidade do sistema, enquanto o mecanismo inspirado na teoria de revisão de crenças representou o principal diferencial da proposta, permitindo identificar e atualizar documentos potencialmente desatualizados e contribuindo para a manutenção contínua da base de conhecimento.

Na avaliação com usuários, o sistema obteve pontuação média de 70,8 no CUQ, nota geral de 8,1 e taxa de recomendação de 85,3%. Os participantes destacaram a facilidade de uso e a interface intuitiva, enquanto as principais limitações estiveram relacionadas à cobertura documental e ao tratamento de erros.

Entre as limitações do estudo, destacam-se a sensibilidade das métricas de similaridade semântica a variações de redação, a natureza não totalmente determinística da cadeia de processamento, a dificuldade de recuperação granular em documentos longos e a ausência de uma comparação direta com arquiteturas RAG convencionais sem o mecanismo de revisão de crenças. De modo geral, os resultados demonstram que a solução proposta é tecnicamente viável, bem aceita pelos usuários e capaz de fornecer respostas fundamentadas em fontes institucionais, evidenciando o potencial da revisão de crenças para apoiar a atualização contínua da base de conhecimento.

6 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho apresentou um sistema conversacional inteligente para interpretação de normativas institucionais do IFPI, integrando modelos de linguagem de grande porte, *Retrieval-Augmented Generation* e uma arquitetura fundamentada em Sistemas Multiagentes, à qual se incorporou um mecanismo inspirado na teoria de Revisão de Crenças. Nesse mecanismo reside a principal contribuição da pesquisa: a adaptação de operações da teoria AGM ao gerenciamento de bases vetoriais, aspecto ainda pouco explorado em soluções conversacionais institucionais baseadas em RAG.

Os objetivos específicos definidos na Introdução foram atingidos, conforme as evidências reunidas na Seção 5:

- **projetar uma arquitetura capaz de recuperar informações institucionais de forma confiável:** sobre um conjunto de 51 questões, o sistema processou a totalidade das consultas sem falhas e fundamentou 78% das respostas em fontes institucionais legítimas, com similaridade semântica média de 0,78; a acurácia semântica de 0,63 e a sobreposição média de fontes de 0,54 delimitam esse alcance quanto à completude factual e à correspondência exata das fontes;

- **implementar um mecanismo de atualização contínua da base de conhecimento:** o mecanismo classificou corretamente os estados de manutenção, atualização e obsolescência nos 65 cenários de verificação executados, resultado que verifica a correta implementação das regras de decisão sobre o domínio para o qual foram calibradas;

- **avaliar a eficácia da solução:** a avaliação com 34 membros da comunidade acadêmica resultou em pontuação média de 70,8 no *Chatbot Usability Questionnaire* (CUQ), nota geral de 8,1 e intenção de recomendação de 85,3%.

Quanto à pergunta central formulada na Introdução, os resultados indicam que a incorporação do mecanismo de revisão de crenças a uma arquitetura baseada em RAG e Sistemas Multiagentes é tecnicamente viável e permite identificar alterações e situações de obsolescência em documentos normativos, refletindo-as na camada de indexação. A resposta é, contudo, parcial: documentos confirmados como obsoletos permanecem recuperáveis pela busca semântica, conforme detalhado na Seção 4.2.5, e a ausência de comparação com uma arquitetura RAG convencional impede mensurar o ganho atribuível especificamente ao mecanismo proposto.

Reconhecidas tais delimitações, a solução mostra-se viável e bem aceita por seus usuários, com potencial para reduzir a sobrecarga do corpo técnico-administrativo e ampliar a autonomia dos usuários no acesso às informações institucionais.

6.1 TRABALHOS FUTUROS

Como trabalhos futuros, propõe-se:

1. a ampliação do conjunto de avaliação técnica, com maior número de perguntas e documentos de outros campi do IFPI, de modo a aumentar a robustez estatística dos resultados;

2. a melhoria na interpretação de consultas complexas e ambíguas, possivelmente com o uso de técnicas de decomposição de perguntas;

3. a evolução do mecanismo de revisão de crenças com abordagens mais avançadas de detecção de mudanças semânticas;

4. a integração com sistemas institucionais já consolidados, como o SUAP, ampliando o escopo de informações acessíveis ao usuário;

5. a realização de estudos comparativos que isolem a contribuição de cada componente da arquitetura, contrastando a abordagem RAG adotada com o envio direto dos documentos no contexto da requisição, com o uso do modelo apoiado apenas em seu conhecimento paramétrico e com uma arquitetura RAG convencional desprovida do mecanismo de revisão de crenças;

6. o desenvolvimento de um mecanismo de arbitragem entre documentos ativos que apresentem informações contraditórias, hoje recuperados simultaneamente por similaridade e conciliados apenas pelo modelo de linguagem;

7. a construção de um conjunto de avaliação específico para ataques de *prompt injection*, aplicado em delineamento fatorial que contraste o sistema sem proteção, com cada camada isoladamente e com ambas em operação, de modo a mensurar a contribuição individual de cada uma.

Por fim, além dos resultados obtidos no contexto do IFPI, o trabalho contribui para a discussão sobre mecanismos de manutenção contínua de conhecimento em sistemas baseados em RAG. A adaptação de conceitos da teoria de revisão de crenças para o gerenciamento de bases vetoriais representa uma abordagem promissora para reduzir problemas de obsolescência informacional, ampliando a confiabilidade de sistemas conversacionais utilizados em ambientes organizacionais dinâmicos.

REFERÊNCIAS

ALCHOURRÓN, Carlos Eduardo; GÄRDENFORS, Peter; MAKINSON, David. On the logic of theory change: Partial meet contraction and revision functions. **Journal of Symbolic Logic**, v. 50, n. 2, p. 510–530, 1985. DOI: 10.2307/2274239.

HOLMBERG, Jan-Erik. Defense-in-Depth. *In:* HANDBOOK of Safety Principles. [*S. l.*]: John Wiley & Sons, Ltd, 2017. cap. 4, p. 42–62. ISBN 9781119443070. DOI: 10.1002/9781119443070.ch4. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119443070.ch4. Acesso em: 23 jul. 2026.

HOLMES, Wendy *et al.* Usability Testing of a Healthcare Chatbot: Can We Use Conventional Methods to Assess Conversational User Interfaces? *In:* MULVENNA, Maurice; BOND, Raymond (ed.). **Proceedings of the 31st European Conference on Cognitive Ergonomics**. New York, NY, USA: ACM, 2019. (ECCE 2019), p. 207–214. DOI: 10.1145/3335082.3335094. Disponível em: https://dl.acm.org/doi/10.1145/3335082.3335094. Acesso em: 23 jul. 2026.

KARPUKHIN, Vladimir *et al.* Dense Passage Retrieval for Open-Domain Question Answering. **CoRR**, abs/2004.04906, 2020. Disponível em: https://arxiv.org/abs/2004.04906. Acesso em: 23 jul. 2026.

LEWIS, Patrick *et al.* Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. **CoRR**, abs/2005.11401, 2020. Disponível em: https://arxiv.org/abs/2005.11401. Acesso em: 23 jul. 2026.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. **Introduction to Information Retrieval**. [*S. l.*]: Cambridge University Press, 2008. ISBN 978-0-521-86571-5.

RUSSELL, Stuart; NORVIG, Peter. **Inteligência Artificial: Uma Abordagem Moderna**. 4. ed. Rio de Janeiro: GEN LTC, 2021.

SAAD, Matthew; QAWAQNEH, Zakariya. Closed Domain Question-Answering Techniques in an Institutional Chatbot. *In:* 2024 4th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME). [*S. l.: s. n.*], 2024. p. 1–8. DOI: 10.1109/ICECCME62383.2024.10796881.

VASWANI, Ashish *et al.* **Attention Is All You Need**. [*S. l.: s. n.*], 2017. Disponível em: https://arxiv.org/abs/1706.03762. Acesso em: 23 jul. 2026.

XIE, Xingrui *et al.* A Brief Survey of Vector Databases. *In:* 2023 9th International Conference on Big Data and Information Analytics (BigDIA). [*S. l.: s. n.*], 2023. p. 364–371. DOI: 10.1109/BigDIA60676.2023.10429609.

ZHAO, Wayne Xin *et al.* **A Survey of Large Language Models**. [*S. l.: s. n.*], 2023. Disponível em: https://arxiv.org/abs/2303.18223. Acesso em: 23 jul. 2026.

ZHU, Mingjie *et al.* A Survey on Large Language Model-based Multi-Agent Systems: Paradigms, Applications, and Challenges. **TechRxiv**, v. 2026, n. 0220, 2026. DOI: 10.36227/techrxiv.177160638.89229642/v1. Disponível em: https://www.techrxiv.org/doi/abs/10.36227/techrxiv.177160638.89229642/v1. Acesso em: 23 jul. 2026.