HOSPITAL ESCOLA DA UNIV. FEDERAL DE PELOTAS

Estudo Técnico Preliminar 56/2026

1. Informações Básicas

Número do processo: 23762.007581/2026-31

2. Descrição da necessidade

O Hospital Escola da Universidade Federal de Pelotas, no âmbito do projeto de pesquisa **Acesso-Fácil-HU**, pretende desenvolver, validar e disponibilizar um sistema de chatbot inteligente destinado a facilitar o acesso dos colaboradores a informações institucionais.

O projeto será baseado em técnicas de **RAG - Recuperação Aumentada por Geração**, utilizando documentos disponibilizados em uma base de conhecimento institucional. Essa arquitetura permitirá que o usuário formule perguntas em linguagem natural e receba respostas elaboradas com base nas informações recuperadas dessa base.

Para viabilizar o funcionamento desse chatbot será necessária a contratação de serviço de acesso, por meio de API, aos modelos de linguagem de grande escala, conhecidos como **LLMs - Large Language Models**. Essa API será utilizada como componente de processamento e geração de respostas da solução, recebendo as perguntas dos usuários e os respectivos contextos recuperados do sistema, processando essas informações e retornando respostas em linguagem natural.

A necessidade da contratação decorre, principalmente, dos seguintes fatores:

- grande volume de documentos institucionais, como manuais, normas, regulamentos, procedimentos operacionais e orientações internas;
- dispersão das informações em diferentes documentos e fontes;
- dificuldade de localização rápida de informações relevantes;
- necessidade de reduzir o tempo empregado em consultas manuais;
- necessidade de disponibilizar uma interface de consulta mais acessível e intuitiva;
- necessidade de apoiar os colaboradores na execução de atividades administrativas, operacionais e assistenciais;
- necessidade de testar e validar a aplicação de modelos de inteligência artificial em ambiente hospitalar-acadêmico;
- necessidade de utilizar modelos com capacidade suficiente para interpretar perguntas, processar o contexto recuperado e produzir respostas coerentes e fundamentadas.

A solução contratada deverá ser integrada ao sistema do Projeto Acesso-Fácil-HU por meio de chamadas programáticas à API. A contratação não abrangerá o desenvolvimento integral do chatbot, a construção da interface de usuários, a criação da base documental, a implementação dos mecanismos de indexação ou recuperação semântica, nem a realização das atividades de divisão e preparação dos documentos em chunks, que serão executadas pela equipe responsável pelo projeto.

Também não se pretende contratar uma plataforma completa de chatbot que incorpore obrigatoriamente todos esses componentes. O objeto necessário limita-se à disponibilização de modelos de linguagem por API, com os recursos técnicos necessários à integração com a arquitetura RAG já prevista no projeto.

A contratação deverá ser realizada com base no consumo efetivo de tokens, abrangendo os tokens de entrada e de saída processados pela API. A estimativa preliminar do projeto é de utilização de aproximadamente **1.000.000.000 (um bilhão) de tokens**, durante o período de execução definido no instrumento contratual.

A contratação de uma API especializada mostra-se mais adequada ao estágio atual do projeto, pois permite:

- acesso a modelos de maior capacidade sem necessidade de aquisição de equipamentos especializados para rodar modelos de LLMs localmente;
- utilização escalável conforme a demanda;
- implantação rápida, redução da complexidade de operação e manutenção;
- pagamento proporcional ao consumo;
- avaliação de diferentes modelos e configurações;
- maior flexibilidade para a pesquisa e para a evolução do chatbot.

A solução deverá ser utilizada como ferramenta de apoio à consulta de informações institucionais. As respostas geradas não substituirão documentos oficiais, normas internas, orientações dos setores responsáveis, decisões administrativas ou avaliações profissionais, especialmente quando envolverem questões clínicas, assistenciais, legais ou de segurança.

A contratação deverá ainda observar os requisitos de segurança da informação, proteção de dados pessoais, controle de acesso, disponibilidade, monitoramento do consumo e não utilização dos dados enviados pelo Hospital para treinamento ou aprimoramento dos modelos.

3. Área requisitante

Área Requisitante	Responsável
Unidade de Sistemas da Informação e Inteligência de Dados	Nathalie Rey da Silva Aquino

4. Necessidades de Negócio

As necessidades de negócio correspondem aos resultados institucionais, operacionais e estratégicos que motivam a contratação da solução, independentemente da tecnologia empregada para sua implementação.

4.1. Apoio ao acesso às informações institucionais

O Hospital Escola da Universidade Federal de Pelotas necessita aprimorar o acesso de seus colaboradores à informações institucionais contidas em manuais, normas, procedimentos, orientações internas, documentos administrativos e demais conteúdos autorizados.

Atualmente, a localização dessas informações pode exigir consultas manuais a documentos extensos e distribuídos em diferentes repositórios. Esse processo pode aumentar o tempo necessário para obtenção das respostas, dificultar a identificação da informação mais adequada e elevar o risco de interpretações incompletas ou equivocadas.

A contratação irá contribuir para tornar esse acesso mais simples, rápido e organizado, permitindo que os usuários formulem perguntas em linguagem natural e recebam respostas baseadas no conteúdo institucional disponibilizado.

4.2. Melhoria da eficiência operacional

A solução irá apoiar a redução do tempo despendido pelos colaboradores na busca, leitura e interpretação de documentos institucionais.

Com o acesso mais ágil às informações, espera-se:

- reduzir o tempo de execução de atividades administrativas e operacionais;
- diminuir a necessidade de consultas repetitivas aos setores responsáveis;
- facilitar a localização de normas, procedimentos e orientações;
- reduzir retrabalho decorrente da busca inadequada de informações;
- melhorar a produtividade dos colaboradores;
- ampliar a autonomia dos usuários na realização de consultas rotineiras.

A solução irá atuar como instrumento de apoio à execução das atividades institucionais, sem substituir os documentos oficiais, os responsáveis pelos processos ou os canais formais de orientação do Hospital.

4.3. Apoio à padronização das orientações

O Hospital necessita facilitar o acesso às orientações institucionais vigentes e reduzir a dispersão das informações entre diferentes documentos e canais.

Ao utilizar uma base documental previamente selecionada e autorizada, o sistema deverá contribuir para:

- promover maior uniformidade nas respostas relacionadas a procedimentos institucionais;
- reduzir interpretações baseadas em informações desatualizadas;
- facilitar a divulgação de normas e orientações;
- apoiar a identificação dos documentos que fundamentam as respostas;
- estimular a utilização dos canais e conteúdos institucionais oficiais.

A responsabilidade pela aprovação, atualização e governança dos documentos continuará pertencendo às áreas competentes do Hospital.

4.4. Apoio às atividades do ambiente hospitalar-acadêmico

O Hospital Escola possui características específicas decorrentes da integração entre assistência, ensino, pesquisa e gestão administrativa. Nesse contexto, diferentes públicos podem necessitar de acesso rápido a informações institucionais, respeitados os perfis de autorização e as responsabilidades de cada usuário.

A solução deverá apoiar, inicialmente, consultas relacionadas a conteúdos institucionais previamente definidos, tais como:

- normas e políticas internas;
- procedimentos administrativos;
- orientações operacionais e manuais de trabalho;
- instruções para utilização de serviços e sistemas;
- informações relacionadas à organização institucional.

A utilização da solução em atividades assistenciais, clínicas ou que envolvam decisões sobre pacientes não integra automaticamente o escopo desta contratação. Qualquer ampliação futura deverá ser precedida de avaliação específica de segurança, privacidade, proteção de dados, responsabilidade profissional e governança.

4.5. Viabilização do Projeto Acesso-Fácil-HU

A contratação é necessária para viabilizar o desenvolvimento, a validação e a disponibilização do Projeto Acesso-Fácil-HU.

O projeto prevê o desenvolvimento de um chatbot baseado em técnicas de Recuperação Aumentada por Geração para permitir consultas a documentos institucionais. A equipe do projeto já está responsável pelo desenvolvimento da aplicação, da arquitetura de recuperação das informações, da organização da base documental e dos demais componentes do chatbot.

Nesse contexto, a necessidade institucional está concentrada na disponibilização do componente responsável pelo processamento das perguntas e pela geração das respostas em linguagem natural.

A contratação irá complementar o desenvolvimento já realizado pela equipe do projeto, evitando a contratação do desenvolvimento integral de um chatbot, consultoria ou serviços de desenvolvimento que resultem em sobreposição de atividades e custos.

4.6. Apoio à inovação e à pesquisa aplicada

O projeto possui natureza de pesquisa aplicada e busca avaliar a utilização de modelos de inteligência artificial em um ambiente hospitalar-acadêmico.

A contratação deverá permitir:

- realizar experimentos controlados com modelos de linguagem;
- comparar configurações e parâmetros de geração;
- avaliar a utilização de técnicas de RAG;
- medir a qualidade das respostas;
- realizar testes com usuários;
- produzir evidências sobre a aplicabilidade da solução;
- apoiar o desenvolvimento de conhecimento institucional em inteligência artificial;
- avaliar riscos, limitações e benefícios do uso da tecnologia.

A solução deverá ser contratada de forma a permitir a avaliação objetiva de seu desempenho, considerando tanto métricas técnicas quanto critérios qualitativos relacionados à utilidade, clareza, coerência e fundamentação das respostas.

4.7. Ampliação da acessibilidade às informações

A solução deverá facilitar o acesso às informações por usuários que não conheçam previamente a estrutura dos repositórios institucionais ou a localização dos documentos.

A possibilidade de formular perguntas em linguagem natural poderá reduzir barreiras relacionadas a:

- conhecimento da estrutura documental;
- utilização de mecanismos tradicionais de busca;
- identificação de palavras-chave;
- interpretação de nomenclaturas administrativas;
- localização de documentos em diferentes ambientes.

A solução deverá apresentar as informações de maneira clara e compatível com o contexto da pergunta, sempre respeitando os limites da base documental disponível e as regras de acesso definidas pelo Hospital.

4.8. Redução da dependência de consultas manuais

A contratação deverá contribuir para reduzir a quantidade de consultas manuais e repetitivas, permitindo que as equipes concentrem seus esforços em atividades de maior complexidade.

A solução não eliminará a necessidade de atendimento humano. Consultas complexas, excepcionais, controversas ou que dependam de interpretação normativa deverão continuar sendo encaminhadas aos setores responsáveis.

4.9. Necessidade de escalabilidade

O Hospital necessita de uma solução que possa ser utilizada progressivamente, acompanhando a evolução do Projeto Acesso-Fácil-HU.

A contratação deverá permitir:

- início controlado em ambiente de desenvolvimento e testes;
- realização de avaliação piloto;
- ampliação gradual do número de usuários;
- expansão do conjunto de documentos autorizados;
- acompanhamento do aumento da demanda;
- controle do consumo conforme a utilização efetiva.

A solução não deverá exigir a contratação antecipada de capacidade superior à demanda estimada, devendo ser possível acompanhar e controlar o consumo ao longo da execução contratual.

4.10. Previsibilidade e controle dos custos

O Hospital necessita de mecanismo de contratação que permita relacionar os custos à utilização efetiva do serviço.

A contratação por volume de tokens deverá ser acompanhada de:

- estimativa de consumo;
- limite máximo contratual;
- monitoramento periódico;
- relatórios de utilização;
- conferência do consumo informado pela contratada;
- controle de custos por período e ambiente;
- medidas de contenção em caso de consumo anormal.

A forma de remuneração deverá possibilitar o planejamento orçamentário e a fiscalização do contrato, evitando cobranças não verificáveis ou baseadas em métricas não relacionadas ao uso efetivo da solução.

4.11. Segurança, privacidade e governança institucional

O Hospital necessita utilizar a solução de maneira compatível com suas obrigações legais e institucionais de segurança da informação e proteção de dados.

A contratação deverá assegurar que:

- os dados enviados à solução não sejam utilizados para treinamento do modelo;
- o tratamento das informações seja limitado à finalidade contratada;
- sejam aplicadas medidas de segurança compatíveis com os riscos;
- sejam evitados o envio e o processamento de dados desnecessários;
- sejam definidos responsáveis pela gestão de usuários e credenciais;
- sejam estabelecidos procedimentos para resposta a incidentes;
- sejam observadas as normas aplicáveis à proteção de dados pessoais.

A utilização inicial deverá priorizar documentos institucionais previamente selecionados e não deverá envolver dados clínicos individualizados ou informações pessoais sensíveis, salvo mediante avaliação e autorização específicas.

4.12. Continuidade e sustentabilidade da solução

O Hospital necessita reduzir os riscos de interrupção do serviço e de dependência excessiva de um único fornecedor ou modelo.

Para isso, a contratação deverá preservar, sempre que possível:

- os dados e documentos sob controle do Hospital;
- os registros de avaliação;
- os parâmetros utilizados;
- os conteúdos de instrução;
- a arquitetura de integração;
- a capacidade de migração para solução equivalente;
- a documentação necessária à continuidade do projeto.

A solução deverá ser integrada à arquitetura desenvolvida pela equipe do projeto de modo a permitir eventual substituição do modelo ou do fornecedor, caso ocorram indisponibilidade, alteração de condições comerciais, perda de desempenho ou outros fatos que comprometam a continuidade do projeto.

4.13. Resultado institucional esperado

Como resultado da contratação, espera-se disponibilizar aos colaboradores do Hospital Escola uma forma mais eficiente, acessível e confiável de consultar informações institucionais, contribuindo para:

- melhoria dos processos internos;
- aumento da produtividade;
- redução do tempo de busca;
- padronização do acesso às orientações;
- fortalecimento da inovação tecnológica;
- desenvolvimento de pesquisa aplicada;
- aprimoramento da experiência dos usuários internos;
- geração de conhecimento para futuras soluções digitais da instituição.

A contratação será considerada alinhada às necessidades de negócio quando possibilitar a utilização efetiva do componente de inteligência artificial do Projeto Acesso-Fácil-HU, com custos controláveis, segurança adequada, disponibilidade compatível e resultados mensuráveis para o Hospital.

5. Necessidades Tecnológicas

A solução a ser contratada deverá disponibilizar, por meio de API, acesso aos modelos de linguagem de grande escala, capazes de interpretar perguntas formuladas em linguagem natural e gerar respostas textuais em português brasileiro.

A API será integrada à aplicação do Projeto Acesso-Fácil-HU, que está sendo desenvolvida pela equipe responsável pelo projeto. Essa integração deverá permitir o envio das perguntas dos usuários, das instruções de funcionamento do chatbot e das informações recuperadas previamente na base documental institucional.

O projeto utilizará uma arquitetura baseada em RAG - Recuperação Aumentada por Geração. Nesse modelo, a aplicação institucional será responsável por organizar os documentos, realizar a divisão dos conteúdos em trechos (*chunks*), executar a recuperação semântica e selecionar as informações relevantes para cada pergunta. A API contratada deverá receber esse contexto e utilizá-lo para elaborar a resposta ao usuário.

Não fazem parte do objeto desta contratação a criação ou organização da base documental, a realização de *chunks*, a geração de *embeddings*, a indexação dos documentos, a recuperação semântica ou o desenvolvimento integral do chatbot. Essas atividades permanecerão sob responsabilidade da equipe do Projeto Acesso-Fácil-HU.

A solução deverá ser compatível com integração por meio de padrões amplamente utilizados, preferencialmente API REST ou tecnologia equivalente, e deverá contar com documentação suficiente para orientar a configuração, a autenticação, o envio das requisições, o recebimento das respostas e o tratamento de erros.

A API deverá permitir o controle dos principais parâmetros de geração de respostas, incluindo, quando disponibilizados pelo modelo, limite de tokens de entrada e saída, temperatura, tamanho máximo do contexto, histórico da conversa, instruções de sistema e identificação do modelo utilizado. Esses recursos são necessários para que a equipe técnica possa ajustar o comportamento do chatbot e realizar testes comparativos entre diferentes configurações.

Os modelos disponibilizados deverão apresentar desempenho adequado para compreender perguntas em português brasileiro e produzir respostas claras, coerentes e compatíveis com o vocabulário institucional e hospitalar. A solução deverá também permitir que o chatbot seja orientado a responder com base nas informações encaminhadas pela aplicação, evitando a criação de conteúdo não fundamentado no contexto fornecido.

Sempre que não houver informação suficiente para responder à pergunta, o modelo será configurado para informar essa limitação ao usuário, em vez de apresentar uma resposta especulativa. Também deverá ser possível estabelecer instruções para que o chatbot não responda a temas fora do escopo definido e encaminhe o usuário aos canais institucionais adequados, quando for necessário.

A comunicação entre a aplicação institucional e a API deverá ocorrer por meio de conexão segura, com utilização de HTTPS/TLS ou tecnologia equivalente. O mecanismo de autenticação deverá

possibilitar o uso de chaves, tokens ou credenciais semelhantes, com possibilidade de revogação ou substituição.

A solução deverá disponibilizar informações sobre os limites técnicos aplicáveis ao serviço, como tamanho máximo do contexto, quantidade máxima de tokens por requisição, limite de requisições, capacidade de processamento simultâneo e políticas aplicáveis quando esses limites forem atingidos. Essas informações são necessárias para que a equipe técnica possa implementar mecanismos de controle, filas, repetição de requisições e tratamento de indisponibilidades.

Também deverão ser disponibilizados mecanismos de acompanhamento do uso da API. A solução deverá permitir, sempre que tecnicamente possível, a consulta à quantidade de requisições realizadas, aos tokens de entrada e saída, ao modelo utilizado, ao período de consumo, aos erros registrados e ao saldo ou limite remanescente.

A contratação deverá admitir medição baseada no consumo efetivo de tokens. A estimativa inicial da demanda é de **1.000.000.000 de tokens (um bilhão de tokens)**, considerando os tokens utilizados nas entradas e nas respostas fornecidas pelo modelo. Essa estimativa não garante o consumo, nem o pagamento, tendo em vista que o pagamento deverá ocorrer com base no consumo efetivo, não no volume de tokens contratado.

A solução deverá apresentar desempenho compatível com o uso interativo previsto para o chatbot, especialmente em relação ao tempo de resposta, à estabilidade e à capacidade de atendimento das requisições. Os níveis de disponibilidade, os procedimentos de manutenção, os canais para comunicação de incidentes e as condições de atendimento deverão ser informados pela contratada e definidos no instrumento contratual.

A plataforma deverá permitir o acompanhamento do consumo e, sempre que disponível, a configuração de limites, alertas, bloqueios ou outras medidas destinadas a evitar utilização indevida ou consumo incompatível com a disponibilidade orçamentária. A aplicação institucional também deverá implementar mecanismos próprios de controle de acesso, limitação de requisições e prevenção contra uso automatizado indevido.

A solução deverá fornecer os dados necessários à fiscalização técnica e contratual, incluindo informações sobre consumo, erros, tempo de resposta, modelo utilizado e período de utilização. A forma de medição e os documentos comprobatórios deverão ser suficientes para permitir a conferência dos valores faturados.

Sempre que possível, deverão ser utilizados padrões de integração que reduzam a dependência da instituição em relação a um modelo ou fornecedor específico. A aplicação, a base documental, as regras de negócio e os mecanismos de recuperação das informações permanecerão sob controle do HE UFPel, de forma a permitir eventual substituição da API com o menor impacto possível.

6. Demais requisitos necessários e suficientes à escolha da solução de TIC

Além dos requisitos tecnológicos, a escolha da solução deverá considerar aspectos relacionados à segurança da informação, à proteção de dados, à continuidade do serviço, ao suporte, à governança, à fiscalização e à viabilidade de manutenção da contratação.

As informações encaminhadas pelo Hospital não deverão ser utilizadas para treinamento ou aperfeiçoamento dos modelos, nem compartilhadas com terceiros para finalidade diversa da execução do serviço. A solução deverá permitir a adoção de medidas de minimização, de modo que sejam enviadas apenas as informações necessárias à elaboração das respostas.

A contratada deverá informar a existência de subcontratados, subprocessadores ou terceiros envolvidos no processamento dos dados, bem como as responsabilidades de cada participante. Também deverá comunicar eventuais alterações relevantes na política de tratamento das informações, na infraestrutura utilizada ou nos procedimentos de segurança.

A solução deverá observar as normas aplicáveis à proteção de dados pessoais, à segurança da informação e à privacidade. Considerando que o serviço será utilizado no ambiente hospitalar, a aplicação deverá ser configurada para evitar o envio desnecessário de dados pessoais ou dados pessoais sensíveis.

O chatbot deverá funcionar como ferramenta de apoio ao acesso a informações institucionais e não deverá ser utilizado para substituir avaliação profissional, decisão clínica, orientação médica, interpretação jurídica definitiva ou determinação administrativa formal.

A contratada deverá adotar medidas de segurança compatíveis com a natureza do serviço, incluindo proteção das credenciais, controle de acesso, criptografia das comunicações, monitoramento de eventos, gestão de vulnerabilidades e procedimentos para comunicação e tratamento de incidentes de segurança.

Deverá ser assegurada a continuidade do serviço durante o período de utilização previsto para o projeto. A contratada deverá informar as condições de disponibilidade da API, as janelas de manutenção programada, os procedimentos aplicáveis em caso de interrupção e os mecanismos de recuperação do serviço.

A equipe do Hospital poderá implementar mensagens de contingência e mecanismos de tratamento de indisponibilidade na aplicação do chatbot, caso necessário.

A solução deverá possuir documentação técnica atualizada e suporte suficiente para permitir sua integração, operação, fiscalização e manutenção. O suporte deverá abranger, no mínimo, esclarecimentos sobre o funcionamento da API, orientação quanto à integração, análise de erros, informações sobre consumo e comunicação de indisponibilidades ou alterações relevantes no serviço.

Os canais de atendimento, os horários de funcionamento e os prazos de resposta deverão ser definidos de acordo com a criticidade da contratação. Também deverão ser estabelecidos procedimentos para abertura, acompanhamento e encerramento de chamados.

Antes da utilização plena, a solução deverá ser submetida a testes e homologação. A avaliação deverá verificar, entre outros aspectos, a autenticidade das credenciais, a comunicação com a API, o envio e recebimento de requisições, o desempenho em língua portuguesa, a compatibilidade com a arquitetura RAG, o consumo de tokens, o comportamento diante de informações insuficientes, o tratamento de erros e os mecanismos de segurança.

A homologação deverá ser realizada com perguntas e situações representativas do uso previsto no Projeto Acesso-Fácil-HU. O objetivo será verificar se a solução consegue produzir respostas coerentes e fundamentadas nas informações encaminhadas pela aplicação.

A contratada deverá comunicar previamente, sempre que possível, mudanças que possam afetar a integração ou o funcionamento do projeto, como descontinuação de modelos, alterações de versão,

mudanças nos limites técnicos, modificações na política de tratamento de dados, alterações de preços ou retirada de funcionalidades. Quando houver descontinuação de modelo ou alteração significativa, deverá ser assegurado prazo razoável para avaliação e adaptação da aplicação.

A solução deverá permitir a identificação do modelo utilizado e, sempre que possível, a rastreabilidade das requisições realizadas. A aplicação institucional deverá manter registros suficientes para apoiar a análise de desempenho, o controle de consumo, a avaliação da qualidade das respostas e a fiscalização contratual, observando os limites legais para armazenamento de dados pessoais.

A contratação deverá estabelecer claramente a divisão de responsabilidades entre o HE UFPel e a contratada. Caberá à contratada disponibilizar e manter a API, os modelos de LLM, a infraestrutura necessária à sua execução, a documentação da API, o suporte para integração da API e os mecanismos de medição do consumo de tokens.

Caberá ao HE UFPel manter a aplicação do chatbot, a base documental, os mecanismos de recuperação das informações, o controle dos usuários, as regras de negócio e os demais componentes que não integram o objeto contratado.

A solução deverá ser escolhida considerando não apenas o preço por token, mas também a qualidade técnica, a adequação ao idioma português, a capacidade de integração com o projeto, a segurança da API, a política de tratamento dos dados, a disponibilidade, o suporte prestado pela contrada, a possibilidade de monitoramento do consumo e a compatibilidade com a arquitetura definida para o Projeto Acesso-Fácil-HU.

Por fim, a solução deverá ser suficientemente flexível para atender ao estágio atual de pesquisa e desenvolvimento do projeto, sem impor a contratação de serviços que não serão utilizados, como construção da base documental, desenvolvimento integral do chatbot, implantação de mecanismos de RAG ou contratação de infraestrutura para armazenamento e processamento dos documentos. A contratação deverá restringir-se ao acesso aos modelos de linguagem por API e aos serviços necessários à sua operação, integração, monitoramento e suporte.

7. Estimativa da demanda - quantidade de bens e serviços

Item	Descrição	Unidade de medição	Quantidade estimada
1	Acesso, por API, a modelos de linguagem de grande escala, incluindo processamento de tokens de entrada e geração de tokens de saída	Token	1.000.000.000 (um bilhão)

A estimativa total é de **1.000.000.000 (um bilhão) de tokens**, considerando o conjunto de tokens utilizados nas requisições enviadas ao modelo e nas respostas geradas pela API.

A quantidade foi estimada para atender às atividades de desenvolvimento, testes, validação e utilização do chatbot no âmbito do Projeto Acesso-Fácil-HU durante o período de vigência contratual. O volume estimado deverá contemplar tanto o consumo decorrente da operação regular quanto o consumo necessário para avaliação de modelos e configurações dos parâmetros de geração de respostas.

7.2. Composição da demanda

O consumo estimado poderá ocorrer nas seguintes etapas:

Etapa	Finalidade
Desenvolvimento e integração	Configuração da comunicação entre o chatbot e a API, testes de autenticação, envio de requisições e recebimento de respostas
Testes técnicos	Avaliação de limites, desempenho, estabilidade, tratamento de erros e compatibilidade com a arquitetura do projeto
Testes funcionais e de qualidade	Comparação de modelos e parâmetros, verificação da compreensão de perguntas em português e avaliação da adequação das respostas
Homologação	Validação da solução com perguntas e contextos representativos do uso institucional
Operação do chatbot	Processamento das perguntas dos usuários e geração das respostas durante a utilização do sistema

A divisão exata do volume entre tokens de entrada e tokens de saída dependerá do modelo utilizado, do tamanho dos contextos recuperados pela aplicação e da quantidade de interações realizadas pelos usuários. Por essa razão, recomenda-se que o instrumento convocatório permita a medição do consumo de forma consolidada, sem que seja necessário diferenciar os tokens de entrada e de saída.

A quantidade de um bilhão de tokens representa uma **estimativa máxima de planejamento**, não constituindo garantia de consumo integral nem obrigação de pagamento por volume não utilizado. O pagamento deverá ocorrer com base no consumo efetivamente realizado e comprovado, conforme os critérios de medição definidos no edital e no contrato.

7.3. Premissas utilizadas na estimativa

A estimativa considera as seguintes premissas:

- o chatbot será integrado à API por meio de requisições de texto;
- a aplicação do Projeto Acesso-Fácil-HU realizará previamente a recuperação das informações relevantes na base documental;
- os trechos recuperados serão encaminhados à API como contexto da pergunta;
- o consumo abrangerá os tokens enviados à API e os tokens gerados nas respostas;
- poderão ser utilizados diferentes modelos ou versões disponibilizados pela plataforma contratada;
- haverá consumo durante as fases de desenvolvimento, testes, homologação e operação;
- o volume efetivo dependerá da quantidade de usuários, do número de perguntas, do tamanho dos contextos e do tamanho das respostas.

7.4. Forma de medição e pagamento

A medição deverá ser realizada com base no número de tokens efetivamente processados pela solução. Devendo ser adotado o modelo de precificação único por token, englobando nessa precificação tanto os tokens de entrada, como os tokens de saída de cada requisição.

A contratada deverá disponibilizar relatórios, que permitam verificar, no mínimo:

- período de consumo (será adotado o pagamento por consumo mensal);
- quantidade de requisições realizadas no mês;
- quantidade de tokens consumidos por requisição (contabilizando tokens de entrada e tokens de saída);
- modelo utilizado;
- eventuais créditos, descontos ou ajustes;
- valor correspondente ao consumo apurado no mês.

8. Levantamento de soluções

8. Levantamento de soluções

Com o objetivo de atender à necessidade do Projeto Acesso-Fácil-HU, foram identificadas as seguintes alternativas de solução:

8.1. Não contratação de serviço de modelo de linguagem

Essa alternativa consistiria em desenvolver o chatbot sem a utilização de uma API de modelo de linguagem ou manter o projeto restrito a uma etapa experimental, sem disponibilização funcional aos usuários.

8.2. Desenvolvimento ou hospedagem de modelo próprio

Essa alternativa consistiria em adquirir, disponibilizar ou contratar infraestrutura computacional própria para hospedar e executar um modelo de linguagem.

A solução poderia envolver servidores com unidades de processamento gráfico, armazenamento, conectividade, mecanismos de segurança, monitoramento e equipe especializada para instalação, operação, atualização e manutenção dos modelos.

8.3. Contratação direta de fornecedor de modelo de linguagem

Nessa modalidade, a Administração contrataria diretamente um fornecedor específico de modelo de linguagem, com acesso ao serviço por API e remuneração baseada no volume de tokens processados.

A solução poderia incluir um único modelo ou um conjunto limitado de modelos disponibilizados pelo próprio fornecedor.

8.4. Contratação de plataforma agregadora ou gateway de modelos

Essa alternativa consiste na contratação de empresa que disponibilize, por meio de uma interface comum de integração, acesso a diferentes modelos de linguagem, próprios ou de terceiros.

A plataforma poderá atuar como intermediária técnica e comercial entre a Administração e os provedores dos modelos, centralizando, conforme as condições da contratação:

- o acesso aos modelos;

- a documentação da API;
- a autenticação;
- o monitoramento do consumo;
- o faturamento;
- o suporte técnico;
- a disponibilização de diferentes modelos ou versões.

8.5. Contratação de projeto completo de chatbot

Essa alternativa consistiria na contratação de empresa para desenvolver e implantar a solução completa, incluindo, conforme o caso:

- desenvolvimento da aplicação;
- construção da interface;
- organização da base documental;
- implementação da arquitetura RAG;
- criação de mecanismos de recuperação;
- integração com modelos de linguagem;
- implantação, suporte e operação do chatbot.

8.6. Contratação de modelo aberto ou solução soberana hospedada por terceiro

Também foi considerada a possibilidade de contratação de empresa que disponibilize modelo aberto, brasileiro ou soberano, hospedado em infraestrutura própria ou em ambiente localizado no Brasil.

Essa alternativa poderá ser especialmente relevante em projetos que envolvam dados pessoais sensíveis, dados clínicos, informações sigilosas ou exigência de processamento em infraestrutura nacional.

No Projeto Acesso-Fácil-HU, contudo, a solução deverá ser avaliada em relação à disponibilidade imediata, ao atendimento dos benchmarks, à maturidade do modelo, à existência de proposta comercial e à compatibilidade com o cronograma da contratação.

9. Análise comparativa de soluções

As alternativas levantadas foram comparadas com base nos seguintes critérios:

- atendimento à necessidade do projeto;
- compatibilidade com a arquitetura RAG;
- possibilidade de integração por API;
- qualidade e desempenho dos modelos;
- flexibilidade para utilização de diferentes modelos;
- necessidade de infraestrutura própria;
- complexidade de implantação e operação;
- segurança e proteção de dados;
- possibilidade de monitoramento e medição do consumo;
- custo para implantação da solução como um todo;
- compatibilidade com os componentes já desenvolvidos;
- prazo necessário para disponibilização da solução.

9.1. Comparação geral

Alternativa	Complexidade de implantação	Infraestrutura própria	Flexibilidade	Compatibilidade com o projeto
Não contratar serviço de LLM	Baixa	Não	Baixa	Inadequada
Desenvolver ou hospedar modelo próprio	Alta	Sim	Alta	Possível, mas inadequada para o momento atual
Contratar fornecedor direto de modelo	Baixa ou média	Não	Baixa ou média	Tecnicamente possível
Contratar plataforma agregadora ou gateway	Baixa ou média	Não	Alta	Mais adequada
Contratar projeto completo de chatbot	Média ou alta	Não necessariamente	Baixa ou média	Inadequada ao escopo definido
Contratar modelo aberto ou soberano hospedado por terceiro	Média	Não necessariamente	Média	Dependente da maturidade e disponibilidade da solução

9.2. Não contratação de serviço de modelo de linguagem

A não contratação não atende à necessidade do Projeto Acesso-Fácil-HU, pois o chatbot depende de um modelo de linguagem para interpretar as perguntas formuladas pelos usuários e gerar respostas com base no contexto recuperado da base documental.

Sem esse componente, seria possível desenvolver a interface, a base de documentos e parte da arquitetura de recuperação, mas não seria possível concluir a solução nem validar adequadamente seu funcionamento.

Essa alternativa prejudicaria:

- a integração entre o chatbot e o modelo de linguagem;
- os testes da arquitetura RAG;
- a avaliação da qualidade das respostas;
- a realização de testes com usuários;

- a validação dos resultados da pesquisa;
- a disponibilização do chatbot em ambiente institucional.

Por esse motivo, a alternativa foi considerada inviável para o atendimento da demanda.

9.3. Desenvolvimento ou hospedagem de modelo próprio

A hospedagem própria proporcionaria maior controle sobre o ambiente de execução e poderia reduzir a dependência de fornecedores externos. Entretanto, exigiria a aquisição ou disponibilização de infraestrutura computacional especializada, além da constituição ou alocação de equipe capaz de operar e manter o ambiente.

A Administração teria de assumir diretamente os custos e riscos relacionados a:

- aquisição ou locação de servidores e GPUs;
- armazenamento e conectividade;
- instalação e configuração dos modelos;
- atualização de bibliotecas e componentes;
- aplicação de correções de segurança;
- monitoramento de capacidade e disponibilidade;
- manutenção, suporte e avaliação contínua do desempenho.

Embora possa ser considerada em projetos futuros, especialmente quando houver necessidade de processamento local ou maior controle sobre dados sensíveis, essa alternativa apresenta complexidade e custos superiores aos necessários para a etapa atual do Projeto Acesso-Fácil-HU.

A hospedagem própria também não asseguraria o atendimento aos patamares mínimos de desempenho definidos para os benchmarks, nem garantiria desempenho adequado em língua portuguesa e na arquitetura RAG.

9.4. Contratação direta de fornecedor de modelo

A contratação direta de um fornecedor específico poderia atender à necessidade técnica do chatbot, desde que o modelo disponibilizado cumprisse os requisitos mínimos de qualidade, integração, segurança, disponibilidade e suporte.

Essa alternativa apresenta como vantagens:

- integração potencialmente simples;
- definição direta das responsabilidades;
- relacionamento direto com o fornecedor do modelo;
- possibilidade de negociação de preços e níveis de serviço;
- menor quantidade de intermediários na cadeia de fornecimento.

Entretanto, a solução tende a gerar maior dependência de uma única plataforma e pode limitar a comparação entre diferentes modelos. Também poderá exigir alterações na aplicação caso o modelo seja descontinuado, tenha seu desempenho alterado ou deixe de atender às necessidades do projeto.

A contratação direta não é tecnicamente inviável. Poderá ser aceita caso a proposta atenda integralmente aos requisitos do edital. Entretanto, considerando o caráter de pesquisa, desenvolvimento e avaliação do Projeto Acesso-Fácil-HU, a solução apresenta menor flexibilidade que uma plataforma com diferentes modelos disponíveis.

9.5. Contratação de plataforma agregadora ou gateway de modelos

A plataforma agregadora ou gateway é a alternativa que apresenta maior aderência às necessidades do Projeto Acesso-Fácil-HU.

Essa solução permite acessar diferentes modelos por meio de uma integração comum, reduzindo o esforço necessário para comparar desempenho, custo, latência, capacidade de geração e adequação ao português brasileiro.

A alternativa também facilita a substituição de modelos, desde que sejam preservados os padrões de integração e os requisitos contratuais. Essa característica é especialmente relevante para um projeto que prevê experimentação, avaliação técnica e possibilidade de evolução da solução.

Os principais benefícios são:

- compatibilidade com a arquitetura RAG;
- integração por API;
- acesso a diferentes modelos;
- redução do esforço de desenvolvimento;
- maior flexibilidade para testes;
- possibilidade de substituição de modelos;
- centralização do controle de consumo;
- contratação e pagamento em moeda nacional;
- redução da dependência de um único fabricante;
- adequação ao modelo de pagamento por tokens.

A contratação deverá exigir transparência sobre a cadeia de fornecimento. A empresa deverá informar quais modelos serão utilizados, quais provedores participam da prestação, onde ocorre o processamento, quais fornecedores podem ter acesso aos dados e quais políticas de retenção e treinamento são aplicáveis.

Atendidos esses requisitos, essa alternativa apresenta o melhor equilíbrio entre flexibilidade, capacidade técnica, prazo de implantação, governança e custo operacional.

9.6. Contratação de projeto completo de chatbot

A contratação de projeto completo poderia ser adequada em situações nas quais a Administração não dispusesse de equipe própria para desenvolver a aplicação, organizar os documentos e implementar a arquitetura RAG.

Essa não é a situação do Projeto Acesso-Fácil-HU. A equipe responsável já está desenvolvendo:

- a aplicação do chatbot;
- a arquitetura do projeto;
- a organização da base documental;
- os mecanismos de recuperação;
- a integração da solução;
- os procedimentos de avaliação.

A contratação de um projeto completo resultaria em sobreposição de atividades e incluiria serviços que não são necessários ao objeto pretendido. Também poderia elevar os custos, reduzir a autonomia técnica da equipe e dificultar a avaliação de diferentes modelos.

Por esse motivo, a contratação deverá ser restrita ao fornecimento da API de modelos de linguagem e aos serviços diretamente relacionados à sua operação.

9.7. Modelo aberto ou solução soberana hospedada por terceiro

A utilização de modelo aberto, brasileiro ou soberano poderá oferecer vantagens relacionadas ao controle tecnológico, à possibilidade de processamento em infraestrutura nacional e à redução da dependência de provedores estrangeiros.

Entretanto, para a presente contratação, essa alternativa depende da existência de solução madura, disponível e capaz de atender cumulativamente aos requisitos definidos, especialmente:

- desempenho mínimo nos benchmarks exigidos;
- janela de contexto;
- capacidade de geração;
- compatibilidade com Function Calling, Structured Outputs e Streaming;
- integração por API;
- desempenho em língua portuguesa;
- disponibilidade mínima;
- suporte técnico;
- apresentação de proposta comercial compatível;
- atendimento ao cronograma do projeto.

Com base nas informações disponíveis durante o levantamento, não foi possível confirmar que essa alternativa atenderia integralmente aos requisitos técnicos e temporais da contratação. Por isso, poderá ser considerada em projetos futuros, mas não foi escolhida como solução principal para o Projeto Acesso-Fácil-HU.

9.8. Conclusão da análise comparativa

Após a comparação das alternativas, conclui-se que a solução mais adequada é a contratação de serviço de acesso a modelos de linguagem por API, preferencialmente por meio de plataforma agregadora ou gateway que disponibilize diferentes modelos em uma interface comum.

Essa solução foi considerada mais vantajosa porque:

- atende diretamente à necessidade do chatbot;
- é compatível com a arquitetura RAG;
- aproveita os componentes já desenvolvidos;
- evita a contratação de desenvolvimento integral da solução;
- dispensa a implantação de infraestrutura própria;
- permite testar e comparar diferentes modelos;
- facilita o monitoramento e a medição do consumo;
- possibilita pagamento proporcional ao uso;
- permite contratação em moeda nacional;
- reduz os efeitos de eventual substituição de modelo ou fornecedor;
- apresenta menor complexidade e menor custo inicial em comparação com as alternativas próprias.

A escolha definitiva deverá ocorrer mediante procedimento de contratação que assegure a competição entre fornecedores aptos a atender aos requisitos estabelecidos. A proposta vencedora deverá comprovar o atendimento aos requisitos técnicos, de desempenho, segurança, proteção de dados, disponibilidade, suporte e medição previstos no edital e no Termo de Referência.

10. Registro de soluções consideradas inviáveis

Durante a análise das alternativas disponíveis, foram consideradas soluções que, embora existentes no mercado ou tecnicamente possíveis, não se mostraram adequadas para atender integralmente às necessidades do Projeto Acesso-Fácil-HU, seja por incompatibilidade com o escopo, elevada complexidade, custo desproporcional ou ausência de requisitos essenciais.

A inviabilidade registrada neste item refere-se à inadequação das alternativas para atender às necessidades e aos requisitos específicos da presente contratação, não significando, necessariamente, que seja tecnicamente inviável ou inaplicável em outros contextos.

10.1. Não contratação de serviço de modelo de linguagem

A não contratação foi considerada inviável porque o chatbot depende de um modelo de linguagem para interpretar as perguntas dos usuários e elaborar respostas com base no contexto recuperado da base documental.

Sem o acesso a uma API de LLM, seria possível desenvolver alguns componentes do projeto, como a interface, a base documental e os mecanismos de recuperação de informações. Entretanto, não seria possível concluir a integração nem validar o funcionamento completo da solução proposta.

Essa alternativa comprometeria:

- o desenvolvimento e a validação do chatbot;
- os testes da arquitetura RAG;
- a comparação entre modelos;
- a avaliação da qualidade das respostas;
- a realização de testes com usuários;
- o cumprimento dos objetivos do Projeto Acesso-Fácil-HU.

Por esse motivo, a não contratação não atende à necessidade institucional.

10.2. Desenvolvimento e operação de modelo próprio

A criação, hospedagem e operação de um modelo próprio foram consideradas inadequadas para o momento atual do projeto.

Essa alternativa exigiria infraestrutura computacional especializada, incluindo servidores de alta capacidade, unidades de processamento gráfico ou recursos equivalentes, armazenamento, mecanismos de segurança, monitoramento e equipe técnica com conhecimentos específicos em inteligência artificial e operação de modelos de linguagem.

Além dos investimentos iniciais, a Administração teria de assumir diretamente atividades como:

- instalação e configuração do modelo;
- atualizações e correções;
- manutenção da infraestrutura;
- monitoramento de desempenho e disponibilidade;
- gerenciamento da capacidade computacional;
- tratamento de falhas;
- implementação de mecanismos de escalabilidade;
- manutenção dos componentes de software;
- avaliação contínua da qualidade das respostas.

A alternativa também não garantiria, por si só, desempenho satisfatório em português brasileiro, compatibilidade com a arquitetura RAG ou qualidade adequada para o contexto institucional. Considerando o escopo atual, o prazo de execução e a infraestrutura disponível, concluiu-se que a solução apresenta complexidade e custo desproporcionais em relação à necessidade identificada.

10.3. Contratação de solução completa de desenvolvimento de chatbot

Foi considerada a contratação de empresa para desenvolver integralmente o chatbot, incluindo a criação da aplicação, organização da base documental, implementação do mecanismo RAG e fornecimento da API de modelo de linguagem.

Essa alternativa foi considerada inadequada porque a equipe responsável pelo Projeto Acesso-Fácil-HU já está desenvolvendo os componentes relacionados à aplicação, à organização das informações e à recuperação dos conteúdos relevantes.

A contratação de uma solução completa poderia resultar em:

- sobreposição de atividades;
- contratação de serviços que não são necessários;
- aumento do custo total;
- redução da autonomia da equipe do projeto;
- maior dependência da contratada;
- dificuldade de separar a parcela efetivamente necessária do objeto;
- limitação da capacidade de pesquisa e comparação de diferentes modelos.

Assim, a contratação deverá se limitar ao serviço de API de modelos de linguagem e aos serviços diretamente necessários à sua utilização, como suporte, documentação, monitoramento e medição do consumo.

10.4. Contratação de infraestrutura própria para execução da solução

Também foi considerada a aquisição ou contratação de infraestrutura própria para hospedar e executar modelos de linguagem.

Essa alternativa foi considerada inviável para a presente demanda em razão da necessidade de investimentos em equipamentos, licenças, armazenamento, conectividade, segurança, manutenção e recursos especializados.

Além disso, a utilização de infraestrutura própria transferiria para a Administração responsabilidades operacionais que não são essenciais ao objetivo do projeto, como a manutenção da disponibilidade da plataforma, o gerenciamento de atualizações e a adequação da capacidade computacional às variações de demanda.

Para o volume estimado de utilização e para a finalidade do Projeto Acesso-Fácil-HU, a contratação de serviço especializado por API apresenta menor complexidade e permite que a equipe se concentre no desenvolvimento e na validação da aplicação institucional.

10.5. Contratação de fornecedor que disponibilize apenas um modelo fixo

A contratação de uma solução limitada a um único modelo fixo foi considerada insuficiente para atender integralmente às necessidades do projeto, especialmente durante as etapas de desenvolvimento, testes e avaliação.

O Projeto Acesso-Fácil-HU possui caráter de desenvolvimento e validação de solução baseada em inteligência artificial. Por isso, pode ser necessário comparar modelos quanto a:

- qualidade das respostas;
- desempenho em língua portuguesa;
- capacidade de seguir instruções;
- compatibilidade com o contexto fornecido pela arquitetura RAG;
- velocidade de resposta;
- custo por token;
- estabilidade e disponibilidade.

A disponibilização de apenas um modelo pode limitar essa avaliação e aumentar a dependência tecnológica em relação a um único produto. Essa alternativa somente poderá ser admitida caso o fornecedor demonstre que o modelo atende integralmente aos requisitos definidos e que a solução oferece condições adequadas de continuidade, suporte, segurança e desempenho.

10.6. Utilização de ferramentas gratuitas ou de uso experimental

Foram consideradas ferramentas gratuitas, versões de demonstração, planos experimentais ou serviços sem garantia formal de disponibilidade.

Essas alternativas foram consideradas inadequadas para uma solução institucional porque, em geral, podem apresentar:

- limites reduzidos de utilização;
- ausência de níveis de serviço;
- indisponibilidade de suporte técnico adequado;
- alterações unilaterais nas condições de uso;
- restrições de integração;
- ausência de relatórios detalhados de consumo;
- limitações de segurança e governança;
- impossibilidade de garantir continuidade durante toda a vigência contratual.

A utilização de tais ferramentas poderia comprometer a continuidade do chatbot, a previsibilidade da execução contratual e a proteção das informações submetidas à API.

10.7. Contratação baseada exclusivamente em pacote fixo de requisições

Também foi considerada a contratação de quantidade fixa de requisições, chamadas ou consultas, sem medição baseada no volume efetivo de tokens processados.

Essa modalidade pode não refletir adequadamente o consumo real, pois diferentes requisições podem apresentar tamanhos significativamente distintos, tanto no conteúdo enviado ao modelo quanto no tamanho das respostas geradas.

Para o Projeto Acesso-Fácil-HU, a medição por tokens é mais adequada porque permite relacionar o pagamento ao volume efetivamente processado pela solução. Também facilita o acompanhamento do consumo, a fiscalização contratual e a projeção de utilização do saldo contratado.

10.8. Conclusão do registro

Com base nas análises realizadas, foram consideradas inadequadas para a presente contratação:

1. a não contratação de serviço de modelo de linguagem;

2. o desenvolvimento e a operação de modelo próprio;
3. a contratação de solução completa de desenvolvimento do chatbot;
4. a aquisição ou manutenção de infraestrutura própria para execução dos modelos;
5. a contratação limitada a um modelo fixo, sem flexibilidade para avaliação;
6. a utilização de ferramentas gratuitas sem garantias institucionais;
7. a contratação baseada exclusivamente em quantidade fixa de requisições.

A solução considerada mais adequada permanece sendo a contratação de serviço de acesso a modelos de linguagem por API, com cobrança baseada no consumo de tokens e, preferencialmente, com possibilidade de utilização de diferentes modelos por meio de uma integração comum. Essa alternativa atende ao escopo efetivamente necessário, aproveita os componentes já desenvolvidos pela equipe da UFPel e reduz a complexidade operacional da contratação.

11. Análise comparativa de custos (TCO)

A análise comparativa de custos foi realizada considerando o **Custo Total de Propriedade** de cada alternativa avaliada, levando em conta não apenas o preço nominal por token ou o valor da proposta comercial.

O TCO compreende todos os custos diretos e indiretos relacionados à contratação, implantação, integração, operação, manutenção, atualização, suporte, fiscalização, continuidade e eventual substituição da solução durante seu ciclo de vida.

11.1. Componentes considerados no TCO

Componente de custo	**Descrição**
Implantação	Configuração inicial, integração, testes, homologação e disponibilização da solução
Infraestrutura	Servidores, GPUs, armazenamento, rede, segurança, energia e demais recursos necessários
Licenciamento e acesso	Licenças, assinaturas, franquias, mensalidades e valores de utilização
Processamento	Tokens de entrada, tokens de saída, requisições ou outra unidade de consumo
Operação	Administração da solução, monitoramento, gestão de usuários e acompanhamento do consumo
Manutenção	Correções, atualizações, ajustes de integração e manutenção dos componentes
Suporte	Atendimento técnico, resolução de incidentes e orientação à equipe do projeto

Segurança e conformidade	Controles de acesso, proteção de dados, auditoria, retenção e tratamento de incidentes
Continuidade	Redundância, contingência, migração, substituição de modelo e preservação da operação
Desmobilização	Encerramento, revogação de credenciais, exportação, eliminação ou migração de dados

11.2. Desenvolvimento ou hospedagem de modelo próprio

A alternativa de desenvolver ou hospedar um modelo próprio apresenta o maior conjunto de custos indiretos e operacionais entre as soluções avaliadas.

Além da eventual aquisição ou contratação de infraestrutura computacional de alto desempenho, seria necessário considerar:

- servidores e unidades de processamento gráfico;
- armazenamento de grande capacidade;
- instalação e configuração do modelo;
- monitoramento da infraestrutura;
- aplicação de correções de segurança;
- contratação ou alocação de profissionais especializados;
- custos de energia, refrigeração e espaço físico, quando aplicáveis.

Essa alternativa também envolveria riscos de obsolescência tecnológica e necessidade de novos investimentos para acompanhar a evolução dos modelos de linguagem.

Embora o preço por requisição pudesse parecer menor após a implantação, o custo total de propriedade seria significativamente superior, especialmente em razão do investimento inicial, dos custos fixos de operação e da necessidade de manutenção especializada.

Dessa forma, a solução própria foi considerada economicamente desvantajosa para o estágio atual do Projeto Acesso-Fácil-HU.

11.3. Contratação direta de fornecedor de modelo

A contratação direta de um fornecedor de modelo de linguagem tende a apresentar custos de implantação e integração relativamente baixos, especialmente quando a API é bem documentada e o modelo atende diretamente às necessidades do projeto.

Os principais componentes de custo seriam o consumo de tokens, suporte técnico prestado pela contratada, custos de integração e monitoramento, além de eventual migração em caso de substituição do modelo.

Essa alternativa poderá apresentar preço unitário competitivo, mas o TCO deverá considerar também o risco de dependência de um único fornecedor. Caso o modelo seja descontinuado, os preços sejam alterados, a disponibilidade seja reduzida ou o desempenho deixe de atender aos requisitos, poderão surgir custos adicionais.

A contratação direta poderá ser economicamente viável em determinadas condições, mas sua vantagem deverá ser analisada considerando não apenas o preço por token, mas também a flexibilidade, a continuidade e o custo de eventual substituição.

11.4. Contratação de plataforma agregadora ou gateway de modelos

Essa modalidade permite acessar diferentes modelos por meio de uma integração comum, realizando o pagamento conforme o volume de tokens efetivamente consumidos. Dessa forma, a instituição não precisa contratar infraestrutura própria, desenvolver integrações independentes para cada modelo ou assumir diretamente a operação dos recursos computacionais necessários.

Embora a plataforma agregadora possa incluir margem comercial ou custos de intermediação, esses valores devem ser analisados em conjunto com as economias obtidas pela ausência de aquisição de infraestrutura, menores custos de operação, possibilidade de utilização de diferentes modelos e utilização conforme a demanda efetiva.

A possibilidade de pagamento por tokens consumidos é especialmente adequada ao projeto, pois evita a contratação antecipada de capacidade fixa e permite que os pagamentos acompanhem a utilização real da solução.

Considerando o custo total, essa alternativa foi escolhida por apresentar a melhor relação entre preço, flexibilidade, capacidade técnica, prazo de implantação e redução de custos indiretos.

11.5. Contratação de projeto completo de chatbot

A contratação de projeto completo de chatbot teria custo superior ao necessário para atender à demanda atual.

Essa solução incluiria serviços como levantamento de requisitos, desenvolvimento da aplicação, organização da base documental, implementação do mecanismo de RAG, integração com a API, implantação, treinamento e suporte da aplicação.

Essas atividades já estão sendo realizadas pela equipe do Projeto Acesso-Fácil-HU. A contratação de um projeto completo provocaria sobreposição de escopo e pagamento por serviços que não são necessários.

Além do valor do acesso ao modelo, a Administração teria de remunerar atividades de desenvolvimento, gerenciamento e implantação que não fazem parte do objeto pretendido.

Também poderiam surgir custos adicionais relacionados à dependência da contratada para alterações, manutenção e evolução do chatbot. Por essa razão, essa alternativa apresenta TCO superior ao da contratação isolada da API e não se mostra economicamente vantajosa.

11.6. Modelo aberto ou solução soberana hospedada por terceiro

A contratação de modelo aberto, brasileiro ou soberano hospedado por terceiro poderia reduzir alguns custos relacionados à dependência de grandes provedores internacionais. Contudo, considerando as soluções atualmente disponíveis e os requisitos definidos para esta contratação, essa alternativa tenderia a apresentar custos superiores.

Também deverá ser considerado o risco de o modelo não atender integralmente aos requisitos de desempenho, janela de contexto, geração de tokens, Function Calling, Structured Outputs, Streaming e disponibilidade. Nesse caso, seriam necessários investimentos adicionais para complementação, adaptação ou substituição da solução.

Embora modelos abertos ou soberanos possam apresentar vantagens estratégicas em futuras contratações, especialmente quando houver necessidade de processamento local no Brasil ou tratamento de dados sensíveis, não se mostraram, neste momento, a alternativa de menor TCO para o Projeto Acesso-Fácil-HU.

11.7. Conclusão da análise de custos

A contratação de plataforma agregadora ou gateway de modelos por API foi considerada a alternativa de menor Custo Total de Propriedade para o Projeto Acesso-Fácil-HU, considerando o conjunto de custos diretos e indiretos envolvidos.

Essa alternativa permite evitar a aquisição ou manutenção de infraestrutura própria, com pagamento conforme o consumo efetivo de tokens processados pela solução, com isso é esperado uma redução nos custos de implantação e manutenção, além de aproveitar o desenvolvimento já realizado pela equipe do projeto.

12. Descrição da solução de TIC a ser contratada

A solução de Tecnologia da Informação e Comunicação a ser contratada consiste na disponibilização de serviço de acesso, por meio de API, a modelos de linguagem de grande escala, para utilização na validação e na operação do chatbot do Projeto Acesso-Fácil-HU.

A contratação deverá fornecer a capacidade de processamento necessária para que o chatbot interprete perguntas formuladas pelos usuários e produza respostas em linguagem natural, utilizando como referência as informações institucionais previamente recuperadas pela aplicação.

A solução será disponibilizada como serviço, sem necessidade de aquisição de equipamentos, instalação de infraestrutura própria ou desenvolvimento de modelo de inteligência artificial pelo HE UFPel.

O acesso ocorrerá por meio de credenciais e mecanismos de autenticação fornecidos pela contratada, mediante integração entre a API e a aplicação desenvolvida para o Projeto Acesso-Fácil-HU.

12.1. Funcionamento da solução

O funcionamento esperado da solução poderá ser resumido da seguinte forma:

1. o usuário formulará uma pergunta no chatbot;
2. a aplicação do Projeto Acesso-Fácil-HU processará a solicitação;
3. o sistema realizará a recuperação das informações relevantes na base;
4. os trechos recuperados serão encaminhados à API como contexto da pergunta;
5. o modelo de linguagem processará a pergunta, as instruções e o contexto fornecido;
6. a API retornará uma resposta textual à aplicação;
7. o chatbot apresentará a resposta ao usuário, conforme as regras e os mecanismos de controle definidos pelo projeto.

A solução deverá ser compatível com essa arquitetura e permitir que a aplicação encaminhe à API, além da pergunta do usuário, as instruções de funcionamento do chatbot e o contexto recuperado da base institucional.

A contratada não será responsável pela organização dos documentos institucionais, pela divisão dos documentos em trechos (*chunks*), pela geração de *embeddings*, pela indexação da base, pela

recuperação semântica, pela construção da interface ou pelo desenvolvimento integral do chatbot. Esses componentes serão implementados e administrados pela equipe responsável pelo Projeto Acesso-Fácil-HU.

12.2. Características gerais do serviço

A solução deverá disponibilizar acesso a um ou mais modelos de linguagem compatíveis com o objetivo do projeto. A possibilidade de utilização de diferentes modelos pode ser útil para permitir a realização de testes comparativos quanto à qualidade das respostas, velocidade de processamento, desempenho em língua portuguesa, compatibilidade com a arquitetura RAG e consumo de tokens.

A API deverá permitir, conforme as funcionalidades disponíveis no modelo ofertado:

- o envio de perguntas, instruções e contexto documental;
- a definição de instruções de sistema;
- a inclusão do histórico da conversação;
- a configuração de parâmetros de geração;
- o recebimento de respostas em texto;
- a identificação do modelo e, quando aplicável, de sua versão;
- a utilização de respostas estruturadas;
- o processamento progressivo da resposta (*streaming*);
- o uso de chamadas de funções (*Function Calling*);
- o tratamento de erros e indisponibilidades, visando auxiliar a identificação e o tratamento de limitações temporárias durante as requisições;
- a configuração de tempos limite e cancelamento de requisições, quando suportados;
- o monitoramento do consumo de cada requisição.

A solução deverá contar com documentação técnica atualizada e suficiente para permitir sua integração, configuração, utilização, monitoramento e fiscalização pela equipe responsável pelo projeto.

12.3. Unidade de contratação e estimativa de consumo

A contratação será realizada com base no volume estimado de **1.000.000.000 (um bilhão) de tokens**, compreendendo os tokens de entrada enviados à API e os tokens de saída gerados pelos modelos, conforme a metodologia de medição adotada pela solução contratada.

O consumo poderá ocorrer durante as etapas de desenvolvimento, integração, testes, avaliação, homologação e utilização do chatbot. A quantidade estimada não representa necessariamente consumo uniforme ao longo da vigência contratual, podendo haver maior utilização em determinados períodos.

O pagamento deverá estar vinculado ao consumo efetivamente realizado, observadas as regras de medição, faturamento e fiscalização estabelecidas no Termo de Referência e no contrato. A contratada deverá disponibilizar mecanismos que permitam acompanhar o consumo por período, modelo utilizado e quantidade de tokens processados.

A quantidade estimada deverá ser considerada como limite de planejamento da contratação, não constituindo garantia de consumo integral pelo HE-UFPel.

12.4. Integração, suporte e operação

A solução deverá ser compatível com a linguagem, os padrões e a arquitetura utilizados no desenvolvimento do Projeto Acesso-Fácil-HU, não sendo necessária a substituição dos componentes tecnológicos já implementados pela equipe do projeto.

A API deverá funcionar por meio de chamadas REST ou tecnologia equivalente e deverá disponibilizar documentação contendo, no mínimo:

- métodos e endereços de acesso;
- formatos de requisição e resposta;
- parâmetros obrigatórios e opcionais;
- mecanismos de autenticação;
- limites de contexto, requisições e tokens;
- códigos e mensagens de erros;
- procedimentos para tratamento de falhas;
- mecanismos de consulta do consumo;
- informações sobre modelos e versões disponibilizadas.

A contratada deverá fornecer suporte técnico relacionado ao funcionamento da API, à autenticação, à integração, ao acompanhamento do consumo e ao tratamento de falhas ou indisponibilidades. Os canais, horários e prazos de atendimento deverão ser definidos no instrumento convocatório e no contrato, de acordo com a criticidade do serviço.

A solução deverá permitir a identificação de falhas de comunicação, erros de autenticação, limites de utilização, indisponibilidades temporárias e demais ocorrências capazes de afetar o funcionamento do chatbot. Sempre que tecnicamente possível, a API deverá retornar códigos ou mensagens que permitam à aplicação realizar o tratamento adequado das falhas.

12.5. Requisitos mínimos de qualidade dos modelos

Os modelos disponibilizados deverão apresentar desempenho compatível com a aplicação de um chatbot baseado em RAG em ambiente hospitalar-acadêmico.

Os requisitos de qualidade foram definidos com base em benchmarks públicos, independentes e auditáveis, aplicáveis de forma isonômica aos fornecedores e sem indicação ou preferência por marca, produto ou modelo específico. Será admitido qualquer modelo que atenda aos patamares mínimos ou apresente desempenho superior.

12.5.1. Benchmarks mínimos de qualidade dos modelos ofertados

Como regra geral, o modelo ofertado deverá atingir cumulativamente:

Benchmark	**Pontuação mínima**
MMLU-Pro	85%
GPQA Diamond	82%

O atendimento deverá ocorrer nos dois benchmarks. Para modelos que já possuam resultados publicados para o MMLU-Pro e para o GPQA Diamond, não será suficiente atingir apenas um dos patamares, o modelo deverá alcançar ou superar simultaneamente as duas pontuações mínimas.

Excepcionalmente, poderá ser aceita a comprovação de apenas um dos benchmarks quando se tratar de modelo comprovadamente recém-lançado que ainda não possua avaliação pública disponível nos dois benchmarks.

Nessa situação, o modelo deverá apresentar resultado igual ou superior ao patamar mínimo no benchmark disponível, seja ele o MMLU-Pro ou GPQA Diamond. Será necessário ainda que o fornecedor faça a demonstração, por documentação oficial ou fonte independente, a data de lançamento e a inexistência de resultado público para o benchmark.

12.5.2. Requisitos técnicos e funcionais mínimos

A solução deverá disponibilizar API de modelo de linguagem de grande escala com capacidade avançada de processamento de linguagem natural e suporte a janela de contexto de, no mínimo, **128.000 (cento e vinte e oito mil) tokens**, permitindo o processamento de documentos extensos.

O modelo deverá suportar a geração de respostas de, no mínimo, **16.000 (dezesseis mil) tokens por requisição**, de modo a permitir a elaboração de respostas extensas quando necessário. Esse requisito corresponde à capacidade de saída que deverá estar disponível no modelo, não implicando a geração dessa quantidade de tokens em todas as respostas.

Nos testes iniciais realizados pela equipe do Projeto Acesso-Fácil-HU, as respostas geradas pelo chatbot apresentaram consumo médio na faixa de **2.000 (dois mil) a 3.000 (três mil) tokens por resposta**, enquanto o contexto encaminhado ao modelo para fundamentar a geração das respostas apresentou consumo médio de aproximadamente **30.000 (trinta mil) tokens por requisição**.

Esses resultados constituem referências preliminares, não representando limites de utilização, nem garantia de manutenção do mesmo perfil em ambiente de produção, uma vez que o consumo poderá variar conforme as perguntas formuladas, o volume de informações recuperadas, as configurações da aplicação e a extensão necessária das respostas.

A solução deverá ser compatível com pipelines de RAG e aceitar contexto externo fornecido pela aplicação, de modo que as respostas possam ser fundamentadas nas informações recuperadas da base documental institucional.

Os modelos deverão gerar respostas em língua portuguesa, em português brasileiro, com grau adequado de coerência, coesão, clareza e aderência ao contexto fornecido, compatível com aplicações corporativas e institucionais.

A API deverá permitir a configuração de parâmetros de geração, tais como temperatura, *top-p*, limite de tokens e outros controles equivalentes disponibilizados pelo modelo.

A solução deverá oferecer suporte a *Function Calling*, *Structured Outputs* com respostas estruturadas em JSON e *Streaming*, por meio de API REST compatível com padrões amplamente utilizados no mercado. A compatibilidade com padrões de integração não deverá ser interpretada como exigência de marca ou fornecedor específico, devendo ser aceitas soluções tecnicamente equivalentes.

A solução deverá permitir a operação com múltiplos usuários ou requisições simultâneas, observados os limites de concorrência e de utilização informados pelo fornecedor, mantendo desempenho e estabilidade compatíveis com a finalidade do projeto.

Também deverá ser possível realizar o monitoramento diário e mensal do consumo, com informações suficientes para o controle técnico, orçamentário e contratual da utilização dos tokens.

12.6. Fundamentação dos patamares de qualidade

Os patamares mínimos foram estabelecidos tomando-se como referência técnica o modelo Google Gemini 2.5, utilizado nos testes iniciais do Projeto Acesso-Fácil-HU. A menção a esse modelo possui finalidade exclusivamente comparativa e não representa indicação de marca, produto ou fornecedor.

Conforme os resultados considerados no planejamento, o referido modelo apresentou desempenho da ordem de 86% no MMLU-Pro e 83% no GPQA Diamond. Os pisos foram fixados em 85% e 82%, respectivamente, de modo a manter nível de qualidade compatível com o projeto, absorver possíveis variações metodológicas e permitir a participação de modelos equivalentes ou superiores.

O MMLU-Pro (Massive Multitask Language Understanding Pro - https://arxiv.org/abs/2406.01574) foi escolhido por avaliar conhecimento e raciocínio em diferentes áreas e níveis avançados de dificuldade, ele foi desenvolvido como uma evolução do benchmark MMLU, tradicionalmente utilizado para avaliar modelos de linguagem em múltiplas áreas do conhecimento.

Segundo o estudo de referência (MMLU-Pro: Um Benchmark de Compreensão de Linguagem Multitarefa Mais Robusto e Desafiador, disponível em https://arxiv.org/abs/2406.01574), o benchmark abrange 14 áreas disciplinares e mais de 12 mil questões, incluindo matemática, física, química, engenharia, direito, economia, psicologia, saúde, computação, biologia, filosofia e história. Essa característica é relevante para uma aplicação institucional, que poderá ser empregada em diferentes temas, unidades administrativas e situações de atendimento ou apoio à decisão assistencial e administrativa.

O GPQA Diamond (Graduate-Level Google-Proof Question Answering Diamond - https://arxiv.org/abs/2311.12022) foi incluído por avaliar raciocínio científico em nível avançado, com questões nas áreas de biologia, química e física. Diferentemente de benchmarks baseados predominantemente em conhecimento geral ou em perguntas facilmente verificáveis por mecanismos de busca, o GPQA procura medir a capacidade do modelo de resolver problemas complexos, nos quais a resposta correta depende de formação especializada, interpretação cuidadosa e encadeamento lógico.

O estabelecimento do patamar mínimo no GPQA Diamond tem por finalidade selecionar modelos com desempenho próximo ou superior ao nível demonstrado por especialistas humanos no subconjunto mais rigoroso do benchmark. O requisito busca reduzir o risco de contratação de modelos que produzam respostas convincentes, porém incorretas, sobretudo em situações que exigem análise aprofundada e nas quais erros de interpretação possam comprometer a qualidade do resultado produzido pelas respostas do ChatBot.

Embora os benchmarks não reproduzam diretamente as tarefas do chatbot, eles fornecem referências públicas para avaliar capacidades gerais de compreensão, conhecimento e raciocínio do modelo.

A comprovação dos benchmarks deverá ser complementada, durante a avaliação da solução fornecida, por testes específicos com perguntas, documentos e contextos representativos do Projeto Acesso-Fácil-HU.

12.7. Fonte de aferição e forma de comprovação

A verificação do atendimento aos patamares dos benchmarks será realizada, preferencialmente, por consulta aos dados públicos consolidados pela plataforma LLMLearner (antiga DataLearner AI), nos seguintes endereços:

- MMLU-Pro: https://llmlearner.com/rankings/mmlu-pro;
- GPQA Diamond: https://llmlearner.com/rankings/gpqa-diamond.

A comprovação deverá identificar claramente o modelo, sua versão, a modalidade de operação avaliada, a pontuação obtida, a data da avaliação e, quando disponível, a metodologia utilizada.

Caso o modelo ofertado não esteja listado na fonte indicada ou os dados estejam indisponíveis, poderá ser aceita documentação técnica oficial ou fonte independente equivalente, desde que permita verificar a pontuação e a correspondência entre o modelo avaliado e o modelo disponibilizado pela contratada.

Para os modelos recém-lançados enquadrados na exceção prevista no item 12.5.1, o fornecedor deverá apresentar:

- identificação precisa do modelo e da versão ofertada;
- comprovação da data de lançamento;
- resultado público disponível em pelo menos um dos benchmarks;
- justificativa ou evidência da inexistência de resultado público no outro benchmark;
- compromisso de informar eventual publicação posterior do resultado faltante, quando solicitado pela Administração.

A Administração poderá realizar consulta complementar ou solicitar documentação adicional quando houver dúvida quanto à identificação do modelo, à versão avaliada, à metodologia utilizada ou à equivalência entre o modelo submetido ao benchmark e o modelo efetivamente disponibilizado na API.

12.8. Disponibilidade, segurança e proteção das informações

A solução deverá possuir disponibilidade mensal mínima de **99,5%**, conforme o Acordo de Nível de Serviço - SLA, definido no Termo de Referência e no contrato. A disponibilidade deverá considerar o período mensal contratado, ressalvadas apenas as exclusões objetiva e previamente estabelecidas no instrumento contratual.

A contratada deverá informar as janelas de manutenção, os canais de suporte, os procedimentos de comunicação de incidentes e a forma de apuração da disponibilidade.

A solução deverá adotar mecanismos adequados de segurança da informação durante a transmissão, o processamento e eventual armazenamento dos dados submetidos à API. A comunicação entre a aplicação institucional e a plataforma deverá ocorrer por canal seguro, mediante HTTPS/TLS ou tecnologia equivalente.

A contratação deverá observar a legislação aplicável à proteção de dados pessoais, à segurança da informação e ao sigilo das informações institucionais. A utilização da API deverá ser limitada aos dados necessários ao funcionamento do chatbot, evitando-se o envio de informações pessoais ou sensíveis.

Deverão ser previstas pelo HE-UFPel ações de orientação e capacitação dos usuários quanto ao uso adequado do chatbot, com ênfase na proteção de dados pessoais e sensíveis, na prevenção do compartilhamento de informações desnecessárias e no uso responsável da ferramenta, observadas as políticas institucionais de segurança da informação e proteção de dados.

As condições relativas ao tratamento, à retenção, ao armazenamento, à utilização e à eventual eliminação dos dados deverão ser apresentadas pela contratada e analisadas pela área responsável. Sendo vedada a utilização dos dados enviados pelo Hospital Escola da UFPel para treinamento, aperfeiçoamento ou ajuste dos modelos fornecidos pela solução contratada.

A contratada deverá informar, quando aplicável, os locais de processamento, os subcontratados ou subprocessadores envolvidos, as medidas de segurança adotadas e os procedimentos para comunicação e tratamento de incidentes.

A solução deverá permitir o controle das credenciais de acesso e, preferencialmente, oferecer recursos para sua criação, revogação, substituição e segregação entre os ambientes de desenvolvimento, testes e produção.

12.9. Resultado esperado

Com a contratação, espera-se disponibilizar ao Projeto Acesso-Fácil-HU o componente de inteligência artificial necessário para que o chatbot receba perguntas em linguagem natural, interprete as solicitações dos usuários, utilize informações recuperadas da base documental institucional e gere respostas coerentes, contextualizadas e adequadas ao escopo definido.

A solução deverá permitir a realização de testes técnicos e funcionais, a avaliação de diferentes modelos e configurações, o monitoramento do consumo e a operação integrada, controlada e mensurável do chatbot.

A contratação deverá restringir-se ao fornecimento de capacidade de processamento de modelos de linguagem por API e aos serviços diretamente necessários à sua operação, integração, suporte, monitoramento, segurança e medição. Não deverão ser incluídos serviços ou componentes já desenvolvidos ou previstos no escopo do Projeto Acesso-Fácil-HU, como a organização da base documental, a realização de *chunks*, a geração de *embeddings*, a recuperação semântica e o desenvolvimento integral da aplicação.

13. Estimativa de custo total da contratação

Valor (R$): 5.920,00

A estimativa de custo total da contratação foi elaborada com base na pesquisa de mercado realizada junto a potenciais fornecedores de modelos de inteligência artificial por API. Entre as cotações obtidas, foi considerada como referência a proposta de menor preço, cuja solução atende aos requisitos técnicos e funcionais estabelecidos para a contratação.

A proposta apresenta **preço unitário único de R$ 5,92 (cinco reais e noventa e dois centavos) por milhão de tokens**, aplicável ao consumo de tokens de entrada e de saída. Considerando a demanda estimada de **1.000.000.000 (um bilhão) de tokens**, equivalente a 1.000 (mil) unidades de um milhão de tokens, o valor total estimado da contratação é de **R$ 5.920,00 (cinco mil, novecentos e vinte reais)**.

O cálculo utilizado para a estimativa é o seguinte: **1.000 unidades de um milhão de tokens × R$ 5,92 = R$ 5.920,00 (cinco mil, novecentos e vinte reais).**

13.1. Adequação da solução utilizada como referência de preço

O modelo utilizado como referência na cotação possui janela de contexto de **1.000.000 (um milhão) de tokens** e capacidade máxima de geração de **128.000 (cento e vinte e oito mil) tokens por requisição**. Essas capacidades superam os requisitos mínimos da contratação, que estabelecem contexto de 128.000 tokens e geração de, no mínimo, 16.000 tokens por requisição.

Conforme as informações fornecidas, trata-se de modelo multimodal, com entrada de texto e imagem e saída em texto, baseado em arquitetura de Mixture of Experts (mistura de especialista), com 320 bilhões de parâmetros totais e 18 bilhões de parâmetros ativos. O modelo possui pesos abertos,

licença MIT e permissão de uso comercial. Essas características são registradas para descrever a solução considerada na pesquisa, não constituindo exigências adicionais de arquitetura, quantidade de parâmetros ou licenciamento para os demais modelos ofertados.

Em relação ao desempenho, o modelo apresenta pontuação de **90,2% no benchmark GPQA Diamond**, superior ao patamar mínimo exigido de 82%. Até o momento considerado na pesquisa, não havia pontuação disponível no MMLU-Pro, em razão de sua disponibilização recente ao mercado, ocorrida em **26 de agosto de 2026**.

Para fins desta estimativa, o atendimento ao requisito de qualidade considera a regra excepcional prevista para modelos recém-lançados que possuam resultado publicado em apenas um dos benchmarks. A aceitação definitiva deverá observar as condições de comprovação estabelecidas no Termo de Referência, incluindo a identificação da versão ofertada, a comprovação da data de lançamento e a verificação da ausência de resultado público no benchmark faltante. Caso estejam disponíveis resultados nos dois benchmarks no momento da avaliação, deverão ser comprovados os patamares mínimos em ambos.

O atendimento aos demais requisitos da solução deverá igualmente ser comprovado no procedimento de contratação, conforme as especificações e os critérios de aceitação definidos no Termo de Referência.

13.2. Condições da estimativa e do pagamento

O valor de **R$ 5.920,00** corresponde à estimativa de despesa para o consumo integral de um bilhão de tokens, segundo o preço unitário da proposta utilizada como referência. Esse valor não representa obrigação de consumo integral nem definição antecipada do fornecedor a ser contratado.

A utilização dessa cotação como referência de preço não implica preferência por modelo, marca ou fornecedor. O procedimento licitatório deverá admitir a participação de interessados que comprovem o atendimento aos requisitos da contratação, sendo a proposta vencedora definida de acordo com os critérios de julgamento previstos no edital.

14. Justificativa técnica da escolha da solução

A solução escolhida consiste na contratação de serviço de acesso, por meio de API, aos modelos de linguagem de grande escala, com utilização integrada ao chatbot do Projeto Acesso-Fácil-HU e remuneração baseada no consumo efetivo de tokens.

A escolha decorre da compatibilidade dessa solução com a arquitetura definida para o projeto. A aplicação desenvolvida pela equipe do projeto será responsável pela interface com os usuários, pela organização da base documental e pela recuperação das informações relevantes. A API contratada será utilizada para processar as perguntas e elaborar respostas a partir do contexto recuperado e encaminhado pela aplicação.

Essa divisão de responsabilidades permite que a contratação se concentre no componente tecnológico que ainda é necessário ao projeto, sem duplicar serviços de desenvolvimento, organização documental ou implantação do mecanismo RAG. A equipe do projeto mantém o controle sobre a aplicação, a base de conhecimento, as regras de negócio e os critérios de utilização do chatbot, enquanto a contratada fornece o acesso aos modelos de linguagem e os serviços necessários à operação da API.

14.1. Adequação à arquitetura do projeto

A API de LLM é tecnicamente adequada porque pode ser integrada à arquitetura RAG prevista para o Projeto Acesso-Fácil-HU. A aplicação encaminhará à API a pergunta do usuário, as instruções definidas para o chatbot e os trechos de documentos recuperados (chunks), que serão utilizados como contexto para a geração da resposta.

A solução contratada não precisará executar a ingestão, a divisão em *chunks*, a geração de *embeddings*, a indexação ou a busca nos documentos. Essa delimitação evita sobreposição com os componentes que já estão sendo desenvolvidos e reduz a complexidade do objeto contratual.

14.2. Acesso a modelos compatíveis com os requisitos de desempenho

A contratação por API permite utilizar modelos de linguagem com capacidade de processamento compatível com a geração de respostas contextualizadas em português. A solução deverá atender aos requisitos mínimos de desempenho definidos no ETP e no Termo de Referência, incluindo a comprovação dos resultados exigidos nos benchmarks MMLU-Pro e GPQA Diamond.

A adoção de benchmarks busca estabelecer critérios objetivos para a comparação das propostas, sem indicar marca, fabricante ou modelo específico. Os patamares mínimos foram definidos com base em referência técnica utilizada nos testes iniciais do projeto, mas deverão admitir qualquer modelo que comprove desempenho equivalente ou superior, nos termos e pelas formas de comprovação estabelecidos no processo.

Os resultados de benchmarks serão complementados por testes específicos com perguntas e documentos representativos do uso pretendido. Essa validação é necessária porque os benchmarks gerais não substituem a avaliação do comportamento do modelo na arquitetura RAG, no idioma português e no contexto institucional do Hospital Escola da UFPel.

14.3. Adequação da contratação por tokens

A remuneração baseada no consumo de tokens é tecnicamente compatível com a forma de utilização do serviço. Cada pergunta encaminhada à API poderá conter quantidade variável de texto de entrada e gerar resposta de extensão variável. A medição por tokens permite relacionar a cobrança ao volume efetivamente processado.

Para garantir a fiscalização, o Termo de Referência deverá definir a unidade de medida em tokens, o procedimento de apuração, o período de medição e os registros que deverão ser disponibilizados pela contratada.

A previsão de até um bilhão de tokens deverá ser tratada como estimativa de demanda, a ser refinada conforme o número de usuários, o volume de interações, o tamanho dos contextos enviados e a extensão média das respostas. O pagamento deverá ocorrer de acordo com o consumo efetivamente comprovado, dentro dos limites contratuais.

14.5. Requisitos de integração e operação

A solução escolhida deverá permitir integração por meio de API REST ou tecnologia equivalente, oferecer documentação técnica e disponibilizar mecanismos de autenticação, controle de consumo e tratamento de erros.

Também deverá informar os limites técnicos aplicáveis, como janela de contexto, capacidade de geração, limites de requisição e condições de uso simultâneo. Esses elementos são necessários para que a equipe técnica possa implementar a integração de modo seguro e previsível.

A disponibilidade mínima, o suporte técnico, a comunicação de alterações e os procedimentos de continuidade deverão ser definidos no instrumento contratual, de forma compatível com a utilização prevista para o chatbot.

14.6. Segurança da informação e proteção de dados

A escolha da solução está condicionada à capacidade de atender aos requisitos institucionais de segurança e proteção de dados. A contratada deverá informar como os dados enviados à API são tratados, por quanto tempo permanecem disponíveis e quais terceiros participam do processamento.

Os dados encaminhados pelo Hospital Escola não deverão ser utilizados para treinamento de modelos. A aplicação deverá enviar somente as informações necessárias à elaboração das respostas e evitar o encaminhamento de dados pessoais ou sensíveis.

Esses requisitos são parte da avaliação técnica da solução, pois a adequação da API depende não apenas da qualidade do modelo, mas também da segurança do tratamento das informações e da transparência da cadeia de processamento.

14.7. Conclusão da justificativa técnica

A contratação de serviço de API de modelos de linguagem é tecnicamente adequada porque fornece o componente necessário à interpretação das perguntas e à geração das respostas do chatbot, integra-se à arquitetura RAG do Projeto Acesso-Fácil-HU e permite que a Equipe do Projeto mantenha sob seu controle os componentes já desenvolvidos.

A solução também favorece a comparação de modelos, a medição do consumo, a gestão dos custos e a evolução do projeto, sem exigir a implantação de infraestrutura própria ou a contratação de serviços de desenvolvimento já previstos no escopo institucional.

Sendo assim, a escolha é justificada pela aderência funcional à necessidade, pela compatibilidade com a arquitetura existente, pela flexibilidade para avaliação de modelos, pela possibilidade de medição objetiva do consumo e pela viabilidade de estabelecer requisitos verificáveis de desempenho, segurança, disponibilidade e suporte.

15. Justificativa econômica da escolha da solução

A escolha da contratação de serviço de acesso a modelos de linguagem por meio de API, com pagamento vinculado ao consumo de tokens, mostra-se economicamente mais adequada ao Projeto Acesso-Fácil-HU quando comparada às alternativas analisadas.

A solução permite que o Hospital utilize a capacidade de processamento de modelos de inteligência artificial sem a necessidade de adquirir, instalar e manter infraestrutura própria especializada. Dessa forma, evitam-se despesas relacionadas à compra ou locação de servidores de alto desempenho, unidades de processamento gráfico, armazenamento, sistemas de refrigeração, conectividade dedicada, licenças, atualização tecnológica, monitoramento e manutenção da infraestrutura.

Também não será necessário desenvolver, treinar ou hospedar um modelo de linguagem próprio. Essas atividades exigiriam investimentos elevados, equipe técnica especializada e prazo de implantação incompatível com as necessidades atuais do projeto. A contratação da API permite utilizar modelos já disponíveis no mercado, reduzindo os custos iniciais e os riscos técnicos e operacionais associados à criação de uma solução própria.

15.1. Adequação do modelo de pagamento por consumo

O pagamento por volume de tokens mostra-se compatível com a natureza da demanda, uma vez que o consumo da solução dependerá da quantidade de perguntas encaminhadas ao chatbot, do contexto enviado ao modelo e do tamanho das respostas geradas.

Esse modelo apresenta as seguintes vantagens econômicas:

- evita o pagamento de uma capacidade fixa que poderá não ser utilizada;
- permite ajustar o consumo ao estágio de utilização do chatbot;
- possibilita maior controle e rastreabilidade das despesas;
- facilita a fiscalização mediante medição do consumo realizado;
- reduz o risco de investimentos antecipados em infraestrutura;
- permite ampliar ou reduzir o uso conforme a necessidade do projeto.

A estimativa de contratação de **1.000.000.000 (um bilhão) de tokens** foi estabelecida para contemplar as etapas de desenvolvimento, integração, testes, avaliação, homologação e utilização do chatbot durante a vigência contratual. O volume estimado deverá ser tratado como limite de planejamento da demanda, com pagamento condicionado ao consumo efetivamente apurado, conforme as regras definidas no Termo de Referência e no contrato.

15.2. Ausência de custos indiretos relevantes para o HE-UFPel

A contratação de uma solução disponibilizada como serviço reduz a necessidade de incorporação de novos custos indiretos ao projeto. A infraestrutura necessária para a execução dos modelos, incluindo os recursos computacionais e a manutenção da plataforma, ficará sob responsabilidade da contratada, observadas as condições estabelecidas no instrumento contratual.

Ao HE UFPel caberá manter os componentes sob sua responsabilidade, especialmente a aplicação do chatbot, a base documental, os mecanismos de recuperação das informações e os controles institucionais de segurança. Como esses componentes já integram o escopo de desenvolvimento do Projeto Acesso-Fácil-HU, a contratação da API evita a duplicação de investimentos e concentra a despesa no recurso efetivamente necessário.

15.3. Resultado da pesquisa de preços e condições de mercado

Foi realizada pesquisa de preços junto ao mercado e a potenciais fornecedores. A pesquisa demonstrou dificuldade para identificar empresas capazes de fornecer, de forma integrada, uma API com acesso a múltiplos modelos de linguagem, compatível com a arquitetura prevista para o Projeto Acesso-Fácil-HU, com contratação baseada em tokens e condições adequadas às necessidades da Administração.

No levantamento realizado, foi identificada uma empresa que apresentou solução compatível com os requisitos técnicos e operacionais definidos para o projeto. O fato de apenas um fornecedor ter apresentado proposta plenamente aderente não deve ser interpretado como comprovação de exclusividade de mercado ou como fundamento automático para contratação direta. A situação será ser formalmente avaliada no processo licitatório, com base na documentação da pesquisa de preços, nas características do mercado e nas manifestações das áreas competentes.

Considerando a orientação pela realização de procedimento licitatório, a proposta do fornecedor que apresentou a solução compatível deverá ser utilizada como referência para a estimativa de preços, sem prejuízo da participação de outros interessados que comprovem atender integralmente aos requisitos estabelecidos no edital.

15.4. Economicidade da solução escolhida

A solução escolhida apresenta economicidade porque combina:

- menor necessidade de investimento inicial;
- ausência de aquisição de infraestrutura especializada;
- redução dos custos de operação e manutenção;
- possibilidade de utilização de diferentes modelos;
- pagamento conforme o consumo;
- maior flexibilidade para testes e evolução do projeto;
- aproveitamento da arquitetura já desenvolvida;
- redução do tempo e dos riscos de implantação;
- possibilidade de controle do consumo e da execução contratual.

A contratação de uma API agregadora ou de uma plataforma que disponibilize diferentes modelos também reduz o risco de dependência imediata de um único modelo específico. Essa característica permite comparar desempenho, qualidade das respostas, consumo de tokens e custo operacional, possibilitando a escolha do modelo mais adequado para cada etapa do projeto, desde que respeitadas as condições previstas na contratação.

15.5. Conclusão da justificativa econômica

Diante das alternativas avaliadas, a contratação de serviço de API de modelos de linguagem, com remuneração baseada no consumo efetivo de tokens, mostra-se a opção mais adequada sob o ponto de vista econômico para atender às necessidades do Projeto Acesso-Fácil-HU.

A solução evita investimentos de maior valor em infraestrutura e desenvolvimento próprio, reduz custos operacionais, permite flexibilidade de uso e vincula a despesa à utilização efetiva do serviço. Essa vantajosidade econômica foi confirmada durante a fase de pesquisa de preços, onde adotou-se critérios objetivos de julgamento e análise das propostas apresentadas pelos fornecedores.

16. Benefícios a serem alcançados com a contratação

A contratação de serviço de API de modelos de linguagem de grande escala permitirá viabilizar a utilização do chatbot do Projeto Acesso-Fácil-HU. A solução deverá proporcionar benefícios institucionais, operacionais, tecnológicos e relacionados à pesquisa e inovação.

16.1. Benefícios institucionais

Espera-se que a contratação contribua para modernizar a forma de acesso dos colaboradores do Hospital Escola da UFPel às informações institucionais.

Por meio do chatbot, os usuários poderão formular perguntas em linguagem natural e receber respostas baseadas nos documentos e conteúdos previamente disponibilizados pela instituição. Isso deverá facilitar a consulta a manuais, normas, procedimentos e orientações internas, sem eliminar a necessidade de observância dos canais e documentos oficiais.

16.2. Benefícios operacionais

A utilização do chatbot deverá reduzir o tempo necessário para localizar e compreender informações relevantes para a execução das atividades do Hospital.

Entre os resultados esperados estão:

- maior agilidade na obtenção de respostas;
- redução do tempo gasto com pesquisas manuais em documentos extensos;
- apoio à execução padronizada de rotinas;
- melhoria da produtividade dos colaboradores;
- possibilidade de atendimento simultâneo a diversos usuários.

A solução deverá funcionar como ferramenta de apoio à decisão e à execução das atividades, não substituindo a análise dos responsáveis pelos processos nem conferindo caráter oficial ou definitivo a respostas que dependam de interpretação administrativa, técnica ou normativa.

16.3. Benefícios relacionados à qualidade da informação

Ao utilizar informações recuperadas da base documental institucional como contexto para geração das respostas, o projeto poderá melhorar a pertinência e a consistência das respostas fornecidas pelo chatbot.

A aplicação deverá ser configurada para orientar o modelo a:

- utilizar prioritariamente as informações institucionais fornecidas no contexto;
- evitar a criação de informações não presentes na base disponibilizada;
- informar quando não houver elementos suficientes para responder;
- indicar a fonte ou o documento relacionado à resposta;
- encaminhar o usuário aos canais competentes, sempre que necessário.

Esses mecanismos poderão reduzir o risco de respostas genéricas ou dissociadas dos procedimentos adotados pelo Hospital, embora não eliminem a necessidade de monitoramento e avaliação contínua das respostas geradas pelo chatbot.

16.4. Benefícios tecnológicos

A contratação permitirá incorporar ao projeto uma capacidade de processamento de linguagem natural sem a necessidade de desenvolver ou hospedar internamente um modelo de linguagem de grande escala.

Com isso, espera-se:

- reduzir o tempo de implantação da solução;
- aproveitar modelos já disponíveis no mercado;
- facilitar a integração com a aplicação em desenvolvimento;
- permitir a avaliação de diferentes modelos e configurações;
- reduzir a necessidade de infraestrutura computacional especializada;
- facilitar a evolução futura do chatbot;
- manter maior flexibilidade para substituição ou alteração do modelo utilizado.

A disponibilização de múltiplos modelos por meio de uma mesma API poderá permitir comparações quanto à qualidade das respostas, ao desempenho, ao consumo de tokens e ao custo operacional.

16.5. Benefícios para pesquisa, inovação e desenvolvimento institucional

Como o Projeto Acesso-Fácil-HU possui finalidade de pesquisa e validação, a contratação também deverá possibilitar a realização de experimentos controlados e a avaliação de diferentes configurações de modelos de linguagem em um ambiente institucional real.

Esse uso poderá gerar conhecimento sobre:

- aplicação de técnicas de RAG em ambiente hospitalar;
- desempenho de modelos de linguagem em português brasileiro;
- qualidade das respostas baseadas em documentos institucionais;
- adequação de diferentes modelos ao perfil dos usuários;
- custos associados ao processamento de consultas;
- requisitos de segurança, governança e supervisão humana;
- limites e riscos do uso de inteligência artificial em atividades institucionais.

Os resultados poderão subsidiar futuras decisões da Universidade e do Hospital sobre a adoção responsável de soluções de inteligência artificial.

16.6. Benefícios econômicos e de gestão

A solução deverá possibilitar o uso de recursos computacionais sob demanda, com pagamento vinculado ao consumo de tokens, evitando investimentos imediatos em infraestrutura própria e reduzindo custos de manutenção, atualização e operação de ambientes especializados.

Também se espera obter maior controle da despesa por meio de:

- acompanhamento do consumo;
- identificação dos modelos utilizados;
- análise da quantidade de tokens processados;
- comparação entre custo e qualidade das respostas;
- definição de limites e mecanismos de controle;
- avaliação periódica da continuidade e da vantajosidade da contratação.

16.7. Indicadores de acompanhamento

Sempre que possível, os benefícios deverão ser acompanhados por indicadores durante a execução contratual e o desenvolvimento do projeto. Poderão ser utilizados, entre outros:

Benefício	Possível indicador
Agilidade no acesso à informação	Tempo médio de resposta e tempo médio para localização da informação
Utilização da solução	Quantidade de consultas realizadas e número de usuários atendidos
Qualidade das respostas	Percentual de respostas consideradas adequadas em avaliações amostrais
Aderência ao conteúdo institucional	Percentual de respostas fundamentadas nos documentos disponibilizados
Disponibilidade do serviço	Percentual de disponibilidade da API
Controle de custos	Consumo de tokens por período e custo médio por consulta

Redução de retrabalho	Quantidade de consultas repetitivas ou encaminhamentos reduzidos, quando mensurável
Segurança e conformidade	Quantidade de incidentes ou ocorrências relacionadas ao uso da solução

Os indicadores deverão ser ajustados conforme a maturidade do projeto, a disponibilidade de dados e os critérios definidos no Termo de Referência.

16.8. Resultado esperado

Ao final da contratação, espera-se que o Projeto Acesso-Fácil-HU disponha de uma solução funcional e integrada à API contratada, capaz de apoiar os colaboradores do Hospital Escola da UFPel na consulta aos documentos institucionais por meio de linguagem natural.

O principal benefício esperado é a melhoria da eficiência, da acessibilidade e da agilidade no acesso ao conhecimento institucional, com utilização controlada de inteligência artificial, supervisão adequada e preservação da responsabilidade da instituição sobre a validade das informações disponibilizadas

17. Providências a serem Adotadas

Antes da contratação, durante a seleção da solução e ao longo da execução contratual, deverão ser adotadas as providências necessárias para assegurar a adequada implantação, utilização, fiscalização e continuidade do serviço.

17.1. Providências anteriores à contratação

A área demandante deverá consolidar os requisitos técnicos, funcionais, de segurança, de proteção de dados, de medição e de fiscalização que deverão constar do Termo de Referência e dos demais documentos da contratação.

Também deverão ser definidos:

- o período de vigência contratual;
- a quantidade estimada de tokens;
- os modelos de linguagem que poderão ser utilizados;
- os critérios de medição e pagamento;
- os limites de consumo e os mecanismos de controle;
- os níveis mínimos de disponibilidade e desempenho;
- os procedimentos para tratamento de falhas e indisponibilidades;
- as responsabilidades da contratada e da equipe do Projeto Acesso-Fácil-HU;
- os requisitos de encerramento contratual e eventual migração para outra solução.

Deverá ser confirmada a existência de disponibilidade orçamentária compatível com a contratação, bem como concluída a pesquisa de preços e realizada a análise de vantajosidade, observando-se os procedimentos aplicáveis à contratação pública.

A equipe responsável deverá ainda validar a arquitetura de integração entre a aplicação do chatbot e a API, incluindo os mecanismos de autenticação, controle de acesso, registro de consumo, tratamento de erros e armazenamento de informações técnicas necessárias à fiscalização.

17.2. Providências relacionadas à segurança da informação e à proteção de dados

Antes do início da utilização da API em ambiente de produção, deverão ser avaliados os riscos relacionados ao envio de informações à plataforma contratada.

A aplicação deverá ser configurada para encaminhar somente os dados necessários ao processamento da pergunta e da resposta, evitando o envio indevido de informações pessoais, sensíveis, sigilosas ou não relacionadas à finalidade do projeto.

Deverão ser estabelecidas regras para:

- classificação das informações que poderão ser enviadas à API;
- anonimização, pseudonimização ou remoção de dados pessoais, quando aplicável;
- controle de acesso às credenciais da API;
- armazenamento seguro das chaves de autenticação;
- registro e monitoramento das requisições;
- prevenção de uso não autorizado;
- tratamento de incidentes de segurança;
- retenção e eliminação dos dados processados.

A solução deverá ser utilizada em conformidade com a legislação aplicável à proteção de dados pessoais, à segurança da informação e ao sigilo das informações institucionais.

Considerando que a ferramenta será utilizada em um ambiente hospitalar, a equipe deverá adotar cautela especial para impedir o envio de dados de pacientes, prontuários, informações clínicas ou outros dados sensíveis. O chatbot deverá operar exclusivamente com informações institucionais previamente selecionadas e autorizadas.

17.3. Preparação da equipe para implantação

A equipe responsável pelo Projeto Acesso-Fácil-HU deverá realizar os procedimentos necessários para integração, configuração e testes da solução.

Deverão ser executadas, no mínimo, as seguintes atividades:

1. criação e configuração das credenciais de acesso;
2. integração da API com o ambiente de desenvolvimento;
3. realização de testes de conectividade e autenticação;
4. validação do envio do contexto recuperado pela aplicação;
5. avaliação do comportamento dos modelos em perguntas representativas;
6. configuração dos parâmetros de geração;
7. implantação de mecanismos de controle de consumo;
8. realização de testes de desempenho, disponibilidade e segurança;
9. homologação da solução;
10. autorização para utilização em ambiente de produção.

A equipe deverá documentar os procedimentos de integração, as configurações adotadas, os modelos avaliados, os resultados dos testes e as ocorrências identificadas durante a implantação.

17.4. Acompanhamento e fiscalização contratual

Deverão ser designados formalmente os responsáveis pela gestão e fiscalização do contrato, observadas as competências institucionais aplicáveis.

A fiscalização deverá acompanhar, entre outros aspectos:

- o consumo de tokens;
- a disponibilidade da API;
- a ocorrência de falhas ou interrupções;
- os tempos de resposta;
- a qualidade e a aderência das respostas;
- o cumprimento dos requisitos de segurança;
- a correspondência entre o consumo medido e o faturamento;
- o atendimento das solicitações e dos níveis de serviço;
- a observância das obrigações previstas no contrato.

O consumo deverá ser acompanhado periodicamente, por meio de relatórios ou ferramenta de monitoramento disponibilizados pela contratada.

Deverão ser estabelecidos mecanismos para identificar consumos anormais, requisições indevidas, uso não autorizado ou aproximação do limite contratado, permitindo a adoção tempestiva de medidas corretivas.

17.5. Continuidade e encerramento

Durante a execução contratual, a equipe deverá manter documentação suficiente para possibilitar a continuidade do serviço, a substituição do modelo ou a migração para outra solução, caso necessário.

Deverão ser preservados, conforme as regras institucionais e de proteção de dados:

- os códigos e configurações utilizados;
- os registros de integração;
- os parâmetros utilizados;
- os resultados das avaliações;
- os dados de consumo;
- a documentação da arquitetura;
- os procedimentos de operação e fiscalização.

Ao final da contratação, deverão ser adotadas as providências necessárias para revogar credenciais, interromper acessos, eliminar ou devolver dados sob responsabilidade da contratada, além de ser avaliado a necessidade de uma nova contratação ou de migração para uma solução alternativa.

18. Declaração de Viabilidade

Esta equipe de planejamento declara **viável** esta contratação.

18.1. Justificativa da Viabilidade

Com base nas informações levantadas e nas análises realizadas neste Estudo Técnico Preliminar, declara-se **viável a contratação de serviço de acesso, por meio de API, a modelos de linguagem de grande escala para atender às necessidades do Projeto Acesso-Fácil-HU e do HE-UFPel.**

A solução mostrou-se adequada porque atende à necessidade de processamento de linguagem natural do chatbot, sendo compatível com a arquitetura baseada em RAG, além de permitir que a instituição mantenha sob sua responsabilidade a organização, a recuperação e o controle dos documentos utilizados como contexto.

A contratação deverá contemplar estimativa de até **1.000.000.000 (um bilhão) de tokens**, observados o período contratual, a disponibilidade orçamentária, os limites de consumo e as condições estabelecidas no Termo de Referência.

A viabilidade da contratação não afasta a necessidade de avaliação contínua dos riscos relacionados à qualidade das respostas, à possibilidade de geração de informações incorretas, à disponibilidade do serviço, à dependência tecnológica, ao tratamento de dados e à evolução dos preços e modelos disponíveis no mercado.

Também deverá ser assegurado que a solução seja utilizada exclusivamente como ferramenta de apoio, sem substituir a responsabilidade dos setores competentes, os documentos oficiais, os procedimentos institucionais ou a análise humana nos casos que exijam interpretação técnica, administrativa, assistencial ou normativa.

Dessa forma, considerando a necessidade apresentada, as alternativas analisadas, a compatibilidade técnica com o projeto, a adequação econômica do modelo de contratação e a possibilidade de adoção de controles de segurança e governança, conclui-se pela **viabilidade técnica, operacional e econômica da contratação**, devendo o processo prosseguir para a elaboração do Termo de Referência e das demais etapas da contratação pública.

19. Responsáveis

Todas as assinaturas eletrônicas seguem o horário oficial de Brasília e fundamentam-se no §3º do Art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

MATHEUS LOPES DE FERNANDES

Membro da comissão de contratação

Assinou eletronicamente em 23/10/2026 às 09:57:09.

ANA PAULA DA SILVA LOURENCON

Membro da comissão de contratação

Assinou eletronicamente em 23/10/2026 às 10:14:15.

NATHALIE REY DA SILVA AQUINO

Membro da comissão de contratação

Assinou eletronicamente em 23/10/2026 às 09:59:09.

MARIANA VARGAS DA CUNHA

Membro da comissão de contratação

Assinou eletronicamente em 23/10/2026 às 10:00:18.

MARCELO DIOGO COLOSSI

Membro da comissão de contratação

Assinou eletronicamente em 23/10/2026 às 10:02:46.

CLAUDEMIR SANTOS DE PAULA

Membro da comissão de contratação

Assinou eletronicamente em 23/10/2026 às 10:32:35.