

Lugares de Memória Digitais: um Sistema Tutor Inteligente Baseado em RAG para a Preservação da História de Pelotas

Lui Gill Aquini¹, Marcelo Clasen Ribeiro², Marcelo da Silva Dias²,
Laura Quevedo Jurgina¹, Marilton Sanchotene de Aguiar¹,
Leomar Soares da Rosa Junior¹,

¹¹Laboratory for Innovation and Research on Educational Solutions (LIRES)
Universidade Federal de Pelotas (UFPel)
Pelotas – RS – Brasil

²Centro de Desenvolvimento Tecnológico – Universidade Federal de Pelotas (UFPel)
Programa de Pós-Graduação em Computação (PPGC)
Pelotas – RS – Brasil

{lgaquini,mcribeiro,mdsdias}@inf.ufpel.edu.br

{lqjurgina,marilton,leomarjr}@inf.ufpel.br

Abstract. *This paper presents a low-cost intelligent tutoring system based on Retrieval-Augmented Generation for teaching and preserving the historical and cultural memory of a municipality in Rio Grande do Sul, Brazil, whose local knowledge is at risk of being lost. The retrieval corpus consists of reference works on local history. The evaluation addressed generation quality, assessed by a domain expert and by an LLM-as-a-Judge approach, and retrieval quality, assessed through analysis of the top-k retrieved passages and a factual verification test using incorrect information. Taken together, these methods point to the potential of this approach for the technological mediation of collective memory in educational and cultural settings.*

Resumo. *Este artigo apresenta um sistema tutor inteligente de baixo custo baseado em Retrieval-Augmented Generation para o ensino e a preservação da memória histórica e cultural de um município do Rio Grande do Sul, cujo conhecimento local corre risco de dissipação. O corpus de recuperação é composto por obras de referência sobre a história local. A avaliação abrangeu a qualidade da geração, por um especialista de domínio e por LLM as a Judge, e a qualidade da recuperação, pela análise dos trechos top-k e por um teste de verificação factual com informações incorretas. Conjuntamente, tais métodos evidenciam o potencial da referida abordagem para a mediação tecnológica da memória coletiva em contextos educacionais e culturais.*

1. Introdução

A preservação dos saberes historicamente ocupa uma posição central na trajetória humana. Mediante a memória coletiva e a transmissão cultural, as comunidades edificam suas identidades, estabelecem vínculos de pertencimento e encontram continuidade entre o passado e o futuro. Sem tal esforço conjunto, fragmenta-se não apenas a história, mas o próprio sentido da origem e da constituição de uma comunidade [de Sousa et al. 2023].

A lembrança não consiste em um ato exclusivamente individual, mas sim em um esforço compartilhado. A memória, sob essa perspectiva, é forjada no interior dos grupos sociais, visto que estes estipulam o que merece ser recordado, a forma da recordação e a duração da preservação de determinado conhecimento [Halbwachs 1997].

Os denominados lugares de memória, tais como cidades históricas, bibliotecas, museus e monumentos, representam pilares na preservação e na transmissão da herança cultural de um povo. Embora frequentemente submetidos a processos de instrumentalização turística ou econômica, os quais por vezes esvaziam seus significados mais profundos, esses espaços mantêm a função essencial de ancorar a identidade coletiva e de preservar o diálogo entre as gerações [Nora and Erll 1997].

O avanço das novas tecnologias, com destaque para a inteligência artificial, reconfigura a maneira pela qual as gerações recentes interagem, aprendem e constroem o conhecimento. Mais do que instrumentos de uso geral, tais inovações introduzem novas abordagens pedagógicas e metodológicas, ampliando as possibilidades de acesso, de personalização e de mediação dos saberes [Akgun and Greenhow 2022]. Esses recursos digitais apresentam potencial para a preservação do conhecimento comunitário, sobretudo em grupos cujos saberes encontram-se sob o risco de dissipação ao longo do tempo. Adicionalmente, promovem o engajamento intrínseco das gerações mais jovens, o qual pode ser transformado em um vetor de aproximação com a memória e com a herança cultural de seus respectivos grupos sociais [Idris et al. 2016].

Diante desse cenário, o presente estudo tem como objetivo o desenvolvimento de um sistema tutor inteligente fundamentado em *Retrieval-Augmented Generation* (RAG), direcionado ao ensino e à preservação da memória histórica e cultural de uma cidade do interior do Rio Grande do Sul. A proposta articula, por conseguinte, os conceitos de memória coletiva, lugares de memória e mediação tecnológica, com o intuito de converter o engajamento das gerações recentes com as tecnologias digitais em uma experiência significativa de reconexão com a identidade e com o patrimônio de sua respectiva comunidade.

Este trabalho encontra-se estruturado da seguinte maneira: a Seção 2 apresenta a fundamentação teórica, a qual aborda os conceitos de memória coletiva, lugares de memória e o papel das tecnologias digitais na preservação cultural; a Seção 3 descreve os trabalhos correlatos; a Seção 4 detalha a metodologia adotada no desenvolvimento do sistema; a Seção 5 exhibe os resultados e a avaliação qualitativa efetuada; por fim, a Seção 6 expõe as considerações finais e as perspectivas para estudos futuros.

2. Referencial Teórico

A popularização e a facilidade de acesso aos *Large Language Models* (LLMs) introduzem novas possibilidades para o desenvolvimento de ferramentas educacionais [Wang et al. 2026]. Fundamentados na arquitetura Transformer [Vaswani et al. 2017] e em mecanismos de atenção que permitem aos modelos capturarem dependências contextuais de longo alcance, os LLMs têm demonstrado notável capacidade de generalização em diversas tarefas. Por sua vez, o modelo BERT consolidou o pré-treinamento em larga escala como uma estratégia central para o desenvolvimento de modelos de linguagem de alto desempenho, viabilizando uma nova geração de aplicações nas áreas

de processamento de linguagem natural, de educação e de recuperação de informação [Devlin et al. 2019].

A metodologia *Artificial Intelligence Education Unplugged* (AIED *Unplugged*) constitui uma abordagem metodológica direcionada à democratização do uso de ferramentas de inteligência artificial na educação em contextos do Sul Global, nos quais as restrições de infraestrutura e de acesso tecnológico mostram-se inerentes à realidade local. Mediante a adoção de modelos de linguagem de código aberto e da técnica de destilação de modelos, a qual consiste na transferência do conhecimento de modelos de grande porte para versões mais compactas e eficientes, torna-se possível reduzir significativamente os custos computacionais operacionais. Tal estratégia viabiliza a implementação de soluções baseadas em LLMs em ambientes com *hardware* restrito, expandindo o alcance dessas tecnologias a populações e a instituições historicamente marginalizadas do acesso à inteligência artificial educacional [Isotani et al. 2023].

A *Retrieval-Augmented Generation* (RAG) consiste em uma abordagem que possibilita a modelos de linguagem pré-treinados acessar e incorporar informações oriundas de bases de dados externas aos seus processos de treinamento, superando uma das limitações centrais dos LLMs: o conhecimento estático, circunscrito ao corte temporal de seus respectivos dados de treinamento. O procedimento opera, tipicamente, por intermédio da recuperação de fragmentos semanticamente relevantes, armazenados em bancos de dados vetoriais, os quais são fornecidos ao modelo como contexto suplementar no instante da geração da resposta [Lewis et al. 2020]. Nessa dinâmica, a qualidade e a confiabilidade do *corpus* de recuperação revelam-se determinantes para a precisão e para a pertinência das respostas elaboradas, aspecto que consolida a curadoria criteriosa das fontes de dados como uma etapa metodológica central em qualquer implementação de RAG, sobretudo em domínios sensíveis, a exemplo das esferas educacional e cultural [Danuarta et al. 2024].

Os sistemas tutores inteligentes consistem em sistemas de conversação autônomos projetados para interagir com estudantes em linguagem natural, geralmente concebidos para atuar em domínios de conhecimento específicos. Diferentemente dos assistentes de propósito geral, essas ferramentas são desenvolvidas com objetivos pedagógicos definidos, tais como o suporte à aprendizagem, a resolução de dúvidas, o fornecimento de *feedback* ou a mediação de conteúdo, característica que exige não apenas precisão nas respostas, mas também adequação ao perfil e ao contexto do público-alvo [Anderson et al. 1985]. Com o avanço dos LLMs e a consolidação de abordagens como a RAG, os sistemas tutores inteligentes passaram a incorporar capacidades de recuperação e de geração de informação mais sofisticadas, ampliando o seu potencial como ferramentas de apoio ao ensino em contextos diversos [Liu et al. 2025].

Estudos sobre a interação de estudantes com *chatbots* educacionais evidenciam que a experiência de uso não se configura como cognitivamente neutra, uma vez que se encontra permeada por respostas emocionais que influenciam diretamente o engajamento e o desempenho nas atividades propostas. Constatou-se que emoções positivas, a exemplo de curiosidade, de satisfação e de entusiasmo, tendem a ampliar o foco e a persistência dos discentes durante as tarefas, ao passo que emoções negativas podem comprometer a motivação e a qualidade da aprendizagem. Esse aspecto corrobora a necessidade de considerar, no *design* de *chatbots* educacionais, não apenas a exatidão técnica das respostas,

mas também a experiência afetiva do usuário enquanto dimensão pedagógica pertinente [Yin et al. 2024].

3. Trabalhos Relacionados

A aplicação de RAG no desenvolvimento de sistemas tutores inteligentes não constitui uma abordagem inédita na literatura. Entretanto, as pesquisas existentes concentram-se predominantemente em contextos de ensino superior [Liu et al. 2025] ou em domínios específicos, a exemplo do ensino de programação [Alario-Hoyos et al. 2024].

Por sua vez, o estudo de Danuarta et al. (2024) investiga a aplicação de RAG no aprimoramento de um *chatbot* educacional preexistente, direcionado a estudantes do ensino básico. O diferencial da pesquisa reside na construção de um conjunto de dados composto por materiais didáticos adequados à faixa etária do público-alvo, evidenciando que a curadoria e a contextualização do *corpus* de recuperação revelam-se determinantes para a qualidade e para a pertinência das respostas geradas [Danuarta et al. 2024].

O artigo de da Costa et al. (2023) apresenta o desenvolvimento de uma aplicação interativa baseada em mesa tangível, instalada em um museu, com o intuito de promover a interação do público com o acervo institucional de maneira física e imersiva. A proposta, concebida especificamente para o contexto do referido espaço, explora o potencial das interfaces tangíveis enquanto mediadoras entre o visitante e o patrimônio cultural, aproximando-se da concepção de lugares de memória ao converter o museu em um ambiente de experiência ativa e de engajamento cultural [da Costa et al. 2023].

A metodologia AIED *Unplugged* foi aplicada em conjunto com LLMs para a avaliação automática da escrita manual de estudantes brasileiros, estabelecendo como princípio norteador a operacionalização da solução de maneira acessível e computacionalmente viável. A referida proposta evidencia o potencial desse paradigma para mitigar limitações de infraestrutura no contexto educacional brasileiro, alinhando-se às diretrizes da presente pesquisa no tocante à democratização do acesso a ferramentas de IA educacional em realidades do Sul Global [Mello et al. 2025].

O estudo de [Danuarta et al. 2024], direcionado ao ensino básico, aproxima-se desta pesquisa devido à atenção conferida ao *corpus* contextualizado, contudo não abrange a dimensão patrimonial ou identitária. De maneira complementar, [da Costa et al. 2023] investiga o engajamento com o patrimônio cultural por intermédio de interfaces físicas imersivas, ressaltando o potencial das tecnologias digitais enquanto mediadoras da memória coletiva.

O estudo de [Mello et al. 2025], por sua vez, emprega LLMs para a execução de tarefas educacionais; contudo, não utiliza mecanismos de recuperação na referida aplicação.

A Tabela 1 sintetiza os trabalhos relacionados identificados na literatura, organizados conforme a tecnologia central empregada, o contexto de aplicação, o público-alvo, a contribuição principal e a relação com este documento. A análise comparativa demonstra que, embora a utilização de RAG em contextos educacionais represente uma tendência ascendente, as iniciativas vigentes concentram-se predominantemente em domínios técnicos ou institucionais específicos, a exemplo do ensino de programação

[Alario-Hoyos et al. 2024] e do suporte acadêmico no ensino superior [Liu et al. 2025], com escassa atenção à preservação das memórias culturais locais.

Tabela 1. Comparativo dos Trabalhos Relacionados

Trabalho	Tecnologia Central	Contexto de Aplicação	Público-Alvo	Contribuição Principal	Relação com este Trabalho
da Costa et al. (2023)	Mesa tangível	Museu	Visitantes em geral	Interface física e imersiva para interação com acervo cultural	Aproximação com lugares de memória e engajamento com patrimônio cultural
Liu et al. (2025)	RAG + Sistema Tutor Inteligente	Ensino Superior	Estudantes universitários	Aplicação de LLMs como sistemas inteligentes no contexto acadêmico	Uso de RAG educacional, porém restrito ao ensino superior
Alario-Hoyos et al. (2024)	RAG + Chatbot	Ensino de Programação	Estudantes de programação	Personalização do suporte ao aprendizado em domínio específico	Uso de RAG educacional, porém em domínio técnico restrito
Danuarta et al. (2024)	RAG + Chatbot	Ensino Básico	Crianças e jovens	Curadoria de dataset com materiais didáticos adequados à faixa etária	RAG educacional com corpus contextualizado
Mello et al. (2025)	LLMs	Avaliação de escrita manual (Brasil)	Estudantes do Ensino Básico	Avaliação automática de escrita manual com foco em acessibilidade computacional	Aplicação do paradigma AIED Unplugged no contexto brasileiro
Este trabalho	RAG + Chatbot	Preservação da memória da comunidade local	Crianças e jovens	Integração da RAG com o contexto da preservação de memória	—

A análise comparativa evidencia que, embora o emprego de RAG em sistemas tutores inteligentes figure na literatura, os trabalhos identificados abordam tais elementos de maneira fragmentada, uma vez que ora exploram RAG ou LLMs em domínios técnicos ou institucionais desvinculados do patrimônio cultural [Liu et al. 2025, Alario-Hoyos et al. 2024, Danuarta et al. 2024, Mello et al. 2025], ora promovem o engajamento do usuário com a memória coletiva mediante interfaces que prescindem de modelos de linguagem [da Costa et al. 2023]. Nenhum dos estudos mapeados concilia, simultaneamente, RAG, preservação da memória cultural e recorte comunitário local.

4. Metodologia

A Figura 1 exibe a arquitetura geral do sistema desenvolvido, estruturada em dois fluxos complementares. O primeiro, referente à indexação, engloba o processamento da fonte de dados textual, a sua segmentação em *chunks*, a geração de *embeddings* e o armazenamento no banco vetorial. O segundo, relativo à inferência, inicia-se pela pergunta do usuário, a qual é codificada como uma *query* e utilizada para recuperar as informações mais pertinentes do banco vetorial. Tais informações são, em seguida, combinadas com as diretrizes do sistema em um *prompt* estruturado, o qual é submetido ao LLM para a geração da resposta final. As subseções subsequentes detalham cada uma dessas etapas, descrevendo as escolhas técnicas e as ferramentas adotadas em cada componente do *pipeline*.

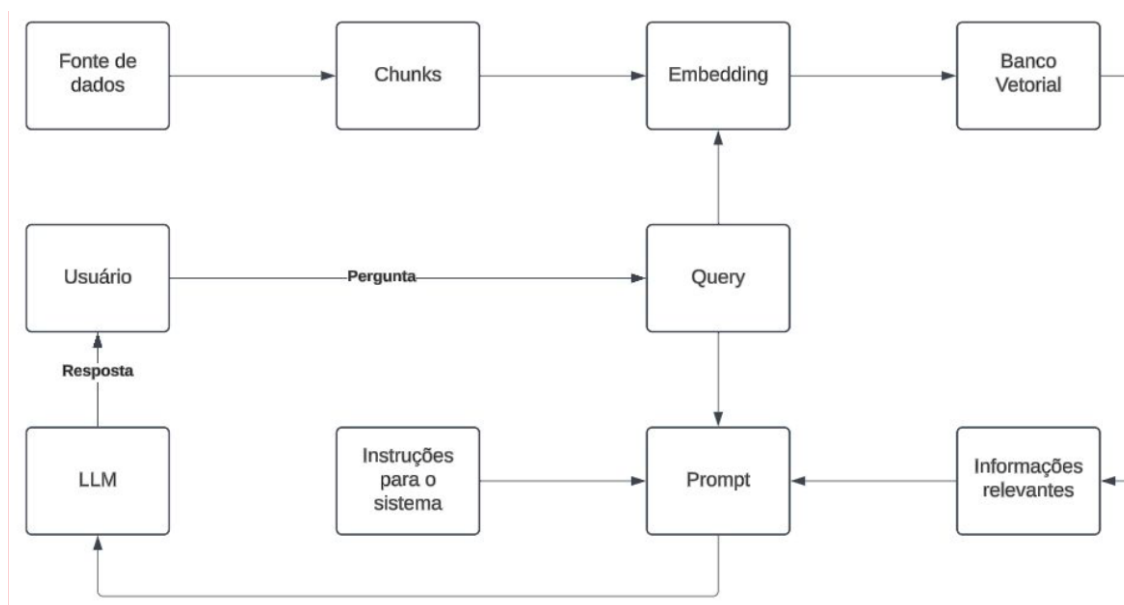


Figura 1. Visão geral do pipeline RAG implementado.

4.1. Hardware Utilizado

Foram utilizadas duas máquinas distintas, mas ambas com o sistema operacional *Ubuntu 26.04 LTS*, no desenvolvimento deste trabalho. A primeira consistia em um *notebook* equipado com processador *Intel Core i5-8350U*, 16 GB de memória RAM e vídeo integrado ao processador. A segunda era um *desktop*, dotado de processador *Intel Core i7-10700K*, 16 GB de memória RAM e placa gráfica dedicada *NVIDIA RTX 2060* com 6 GB de VRAM. Os experimentos envolvendo LLMs foram conduzidos em ambas as máquinas, com ênfase na aferição do tempo de inferência dos modelos de linguagem locais, comparando-se um ambiente provido de placa de vídeo dedicada com outro que dispunha apenas de vídeo integrado.

4.2. Corpus e Processamento de Texto

Como *corpus* de recuperação, selecionaram-se três livros de referência acerca da história e da memória cultural do município de Pelotas, englobando obras de ampla circulação e de reconhecimento regional. A curadoria de tais fontes foi conduzida como uma etapa metodológica central, uma vez que a qualidade do *corpus* de recuperação revela-se determinante para a precisão e para a pertinência das respostas geradas pelo sistema.

No tocante à obtenção do *corpus*, os livros de referência foram adquiridos diretamente com a autora das obras, fator que confere ao projeto não apenas respaldo legal para a utilização do material, mas também uma dimensão colaborativa relevante, aproximando a pesquisa acadêmica, proveniente de diferentes áreas do conhecimento, da comunidade local e dos próprios guardiões da memória cultural que o estudo almeja preservar.

O pré-processamento do *corpus* iniciou-se com a conversão dos documentos PDF para Markdown, mediante o emprego da ferramenta *open-source* *MarkItDown*, desenvolvida pela Microsoft [Microsoft 2024]. No texto convertido, executou-se uma etapa de limpeza, consistente na remoção de quebras de linha espúrias, as quais representam artefatos recorrentes no processo de conversão de PDFs digitalizados. Essa higienização mostra-

se necessária a fim de assegurar a coerência das sentenças previamente à segmentação, impedindo que tais artefatos comprometam a qualidade dos *chunks* gerados.

Na sequência, o texto higienizado foi submetido a uma estratégia de *chunking* fundamentada na segmentação por sentenças, adotando-se o reconhecimento da pontuação final como critério delimitador. O tamanho máximo de cada *chunk* foi estabelecido em 600 caracteres, parâmetro selecionado com o intuito de equilibrar a granularidade semântica e a quantidade de contexto fornecida ao modelo no momento da recuperação. As sentenças que individualmente excedam 1000 caracteres são subdivididas em fragmentos menores previamente à sua incorporação ao *chunk* corrente, assegurando que nenhum *chunk* transponha o limite estipulado.

4.3. Geração de *Embeddings* e Banco Vetorial

Para a geração dos *embeddings*, cada *chunk* elaborado na etapa de processamento de texto foi codificado pelo modelo qwen3-embedding:0.6b [Zhang et al. 2025], disponibilizado nativamente no Ollama, aspecto que suprime a necessidade de configurações suplementares ou de dependências externas para a sua execução. Tal integração nativa com o Ollama revelou-se determinante para a seleção do modelo, visto que possibilita que tanto a geração de *embeddings* quanto a inferência do LLM operem no interior do mesmo ambiente local, simplificando a arquitetura do sistema. O qwen3-embedding:0.6b constitui um modelo de código aberto com desempenho competitivo em tarefas de recuperação semântica, sendo amplamente empregado em *pipelines* RAG de baixo custo operacional.

Para o armazenamento e a recuperação dos documentos, empregou-se o ChromaDB [Chroma 2022] como banco de dados vetorial. Tal escolha justifica-se por sua leveza, por sua facilidade de integração com *pipelines* em Python e por sua compatibilidade com ambientes de execução local, preservando a coerência com os princípios de acessibilidade computacional que norteiam o projeto. A recuperação é conduzida com base na métrica de similaridade padrão do ChromaDB, a *Squared L2 Norm*, retornando os cinco *chunks* mais pertinentes a cada consulta para a composição do contexto fornecido ao LLM.

4.4. Recuperação e Construção do *Prompt*

Ao receber a pergunta do usuário, o sistema a codifica mediante o mesmo modelo de *embeddings* empregado na indexação e executa uma busca por similaridade no ChromaDB, recuperando os cinco *chunks* semanticamente mais próximos à consulta. Esses fragmentos, extraídos diretamente das obras históricas que compõem o *corpus*, funcionam como o substrato documental da resposta. É por intermédio deles que o sistema acessa e mobiliza a memória registrada nos livros de maneira acessível, consultável e mediada pela linguagem natural. Os *chunks* recuperados são, por conseguinte, combinados com as diretrizes de tom e de estilo, carregadas externamente via arquivo YAML, compondo o *prompt* final submetido ao LLM.

Tais diretrizes foram elaboradas com fundamento em um cenário de contação de histórias, orientando o modelo a expor o conteúdo histórico de maneira narrativa, de modo que cada resposta gerada não se restrinja à recuperação de informações, mas contribua para a construção de novos lugares de memória [Nora and Erll 1997] na experiência do usuário.

4.5. Geração de Resposta

Para a implementação do sistema tutor inteligente, adotou-se o Ollama [Ollama 2023] como plataforma de execução local de modelos de linguagem, em consonância com os princípios do paradigma AIED *Unplugged* no tocante à acessibilidade computacional e à redução de custos operacionais. O Ollama expõe uma *Application Programming Interface* (API) local, a qual possibilita a integração direta com o *pipeline* RAG ao receber o *prompt* estruturado, composto pelas instruções do sistema e pelos *chunks* recuperados, e ao retornar a resposta gerada pelo modelo.

Como modelo de linguagem, empregou-se o Gemma3:4b 3B [Team et al. 2025], um modelo de código aberto de pequeno porte que viabiliza a execução em *hardware* convencional, que é acessível para o público em geral, sem perda significativa de desempenho para o domínio proposto. A geração é executada com os parâmetros padrão do modelo, sem ajustes adicionais de temperatura ou de amostragem, opção justificada pelo escopo exploratório do trabalho e pela natureza factual do conteúdo histórico, contexto para o qual respostas mais determinísticas revelam-se mais adequadas. Cada interação é tratada de maneira independente, sem a manutenção de um histórico conversacional, aspecto que simplifica a arquitetura e reduz o consumo de memória em ambientes de uso público.

A interface do sistema foi desenvolvida com o *framework* Streamlit [Streamlit 2019], o que proporciona uma experiência de uso simples e acessível, sem exigir conhecimento técnico por parte do usuário final. A interação ocorre por intermédio de uma caixa de texto, na qual o usuário submete a sua pergunta em linguagem natural e recebe a resposta diretamente na tela. Tal simplicidade constitui uma diretriz intencional, dado que o sistema foi concebido para aplicação no contexto de um museu, ambiente no qual o público-alvo apresenta-se diversificado e o engajamento com a interface deve ocorrer de forma imediata e intuitiva, desprovido de barreiras de usabilidade. A Figura 2 apresenta o protótipo desenvolvido para a primeira versão da aplicação.

5. Resultados

A avaliação do sistema foi estruturada em duas etapas distintas, alinhadas aos dois componentes que compõem a arquitetura RAG: a recuperação e a geração. Na primeira etapa, analisou-se o desempenho do ChromaDB na seleção dos trechos do *corpus* mais pertinentes a cada consulta, mediante a inspeção das distâncias retornadas e de um experimento controlado de injeção de informações no *corpus*. Na segunda etapa, avaliou-se a qualidade das respostas produzidas pelo modelo generativo a partir dos trechos recuperados, ponderando dimensões como a fidelidade ao *corpus*, a cobertura informacional, a adequação contextual e a fluência. Tal segregação possibilitou identificar, de maneira isolada, a origem das principais limitações constatadas, contribuindo para um diagnóstico mais acurado do comportamento do sistema e nortear as diretrizes de aprimoramento discutidas adiante.

A avaliação do componente de recuperação foi conduzida por intermédio de duas abordagens complementares. A primeira consistiu na análise quantitativa dos *top-k chunks* retornados pelo ChromaDB, empregando-se a distância euclidiana quadrática (*squared L2 norm*) entre o vetor de consulta e os vetores dos trechos indexados como métrica de similaridade. Essa análise permitiu aferir o grau de proximidade entre a per-



Figura 2. Protótipo de interface desenvolvida.

gunta do usuário e os fragmentos do *corpus* efetivamente fornecidos como contexto ao modelo gerador.

A segunda abordagem consistiu em um experimento controlado de injeção de informações deliberadamente incorretas no *corpus*. Com o intuito de isolar o efeito da etapa de recuperação, as informações fabricadas foram inseridas em um documento auxiliar, mantido separadamente dos volumes principais utilizados como fonte primária. Tal separação assegura que, caso o sistema reproduza as referidas informações em uma resposta, isso apenas possa ocorrer mediante a recuperação bem-sucedida do documento auxiliar pelo ChromaDB, e não por contaminação implícita do contexto geral ou por conhecimento paramétrico do modelo. Os resultados confirmaram esse comportamento, uma vez que, ao serem formuladas perguntas direcionadas ao conteúdo fabricado, o sistema recuperou o trecho correspondente e reproduziu a informação incorreta na resposta. Esse fato evidencia que a geração encontra-se efetivamente condicionada ao conteúdo retornado pela etapa de recuperação, e não ao conhecimento prévio do modelo de linguagem.

Para a avaliação do sistema de geração, adotou-se uma abordagem fundamentada na comparação entre as respostas geradas pelo *chatbot* e as respostas fornecidas pela própria autora das obras utilizadas como *corpus*, bem como na utilização de um LLM de alta capacidade como uma segunda opinião, metodologia conhecida como *LLM as a Judge* [Zheng et al. 2023]. Um conjunto de dez perguntas foi submetido simultaneamente ao modelo e à autora, cuja *expertise* acerca do conteúdo das obras a posiciona como referência privilegiada para a avaliação da precisão e da pertinência das respostas geradas. O sistema de *LLM as a Judge* foi desenvolvido após a geração, empregando as mesmas perguntas e respostas da primeira etapa.

A comparação entre as respostas geradas pelo sistema e as fornecidas pela autora das obras revelou um alinhamento satisfatório nos aspectos descritivos e conceituais, indicando que o sistema foi capaz de recuperar e de sintetizar uma parcela relevante do *corpus* de maneira fiel ao conteúdo original. Observou-se, contudo, uma fragilidade recorrente no tratamento de informações cronológicas: datas de fundação, períodos de mandato e sequências históricas apresentaram imprecisões em parte das respostas, incluindo casos nos quais o modelo gerou intervalos temporais inconsistentes com o registro contido no *corpus*. Essa limitação mostra-se coerente com o comportamento constatado em modelos generativos de menor porte, os quais tendem a exibir uma menor robustez em tarefas que demandam precisão numérica e ordenação temporal estrita, sugerindo que aplicações futuras, direcionadas à preservação da memória histórica, devam incorporar mecanismos complementares de verificação cronológica [Bhatia et al. 2025].

Para a abordagem de LLM *as a judge*, empregou-se o modelo Claude Opus 4.7, desenvolvido pela Anthropic, ao qual foi fornecido um *prompt* estruturado com o intuito de classificar as respostas do *chatbot* em quatro dimensões: fidelidade ao *corpus*, cobertura informacional, adequação contextual e fluência sobre o assunto. Para cada par de pergunta e resposta, o avaliador atribuiu uma pontuação inteira de 1 a 5 em cada dimensão, acompanhada de uma justificativa textual e da identificação do modo de falha dominante.

A adoção de modelos de linguagem executados localmente por intermédio do Ollama demonstrou resultados positivos no que concerne à privacidade e à soberania da aplicação, transferindo o controle integral do sistema para a própria instituição responsável pela guarda e pela mediação do patrimônio cultural em questão. Tal característica apresenta-se especialmente relevante em contextos institucionais sensíveis, nos quais o envio de dados a servidores externos pode representar tanto uma vulnerabilidade quanto uma dependência tecnológica indesejada.

A avaliação automatizada conduzida pelo Claude Opus 4.7 apresentou um padrão heterogêneo de desempenho, com fluência média elevada (4,7) e fidelidade satisfatória ao *corpus* (4,2), porém com cobertura informacional substancialmente inferior (2,0), o que indica que o principal gargalo do sistema reside na etapa de recuperação, e não na de geração. Perguntas factuais e diretas, cujas respostas encontram-se concentradas em um único trecho do *corpus*, apresentaram resultados positivos, ao passo que questionamentos mais abrangentes ou que demandavam a integração de múltiplos *chunks* evidenciaram as principais limitações do sistema. Nesses cenários, a curadoria criteriosa das fontes e a engenharia de *prompt* revelaram-se suficientes para assegurar respostas adequadas ao perfil do público-alvo. Adicionalmente, verificou-se uma elevação significativa no tempo de inferência em *hardware* de menor capacidade computacional, especificamente pela falta de uma placa gráfica dedicada, notadamente nas consultas que exigiam a síntese de informações dispersas.

6. Considerações Finais

O presente trabalho apresentou o desenvolvimento de um sistema tutor inteligente baseado em RAG, direcionado ao ensino e à preservação da memória histórica e cultural de um município do interior do Rio Grande do Sul. A iniciativa articula os conceitos de memória coletiva e de lugares de memória ao potencial das tecnologias digitais para

o engajamento das gerações mais jovens com o patrimônio de sua comunidade. A proposta insere-se em uma interseção ainda escassamente explorada na literatura, conforme evidenciado pela análise dos trabalhos relacionados, a qual revelou a ausência de iniciativas que combinem, de maneira integrada, RAG educacional e preservação da memória cultural em contextos comunitários locais.

Os resultados obtidos revelam-se promissores, notadamente no que concerne ao desempenho do sistema em perguntas de menor complexidade, cenários nos quais o sistema demonstrou uma satisfatória capacidade de recuperação semântica e de geração de respostas precisas e adequadas ao público-alvo. A adoção de ferramentas de código aberto, a exemplo de Ollama, ChromaDB, Gemma3:4b e qwen3-embedding:0.6b, em consonância com os princípios do paradigma AIED *Unplugged*, demonstrou constituir uma estratégia viável para o desenvolvimento de soluções de IA educacional acessíveis, de reduzido custo operacional e sob o controle institucional do próprio museu, isentas de dependência de infraestrutura externa.

Contudo, o estudo também evidenciou limitações que demandam atenção em investigações futuras. A análise das respostas apontou que o principal gargalo do sistema reside na etapa de recuperação. Em uma parcela significativa das perguntas avaliadas, o conteúdo necessário encontrava-se presente no *corpus*, porém não foi recuperado pelo ChromaDB. Observaram-se, igualmente, imprecisões no tratamento de datas e de sequências históricas, além da elevação do tempo de inferência em *hardware* desprovido de aceleração por placas de vídeo compatíveis com o Ollama.

Como trabalhos futuros, propõe-se o teste de estratégias de *chunking* semântico e de modelos de *embedding* de maior capacidade, a exemplo das variantes mais robustas da família Qwen3-Embedding, com o intuito de mitigar o problema de recuperação. A quantização de modelos e a expansão do *corpus* a partir da inclusão de novas obras de referência também constituem direcionamentos relevantes, contribuindo, respectivamente, para a eficiência computacional e para a cobertura temática do sistema na mediação da memória cultural de Pelotas. Pretende-se, ainda, implementar a ferramenta em museus e ambientes escolares, permitindo a interação direta dos estudantes e consolidando seu uso como recurso educacional.

Declaração sobre uso de Inteligência Artificial

Durante a elaboração deste artigo, ferramentas de inteligência artificial foram utilizadas de forma pontual como apoio à escrita acadêmica, especialmente para revisão ortográfica e de pontuação, além de auxílio na organização de tabelas e na correção de sintaxe em $\text{\LaTeX}$. A geração, análise e interpretação dos dados, bem como as conclusões apresentadas, são de responsabilidade exclusiva dos autores.

Agradecimentos

Agradecemos à FAPERGS, ao CNPq e à CAPES pelo fomento à pesquisa, sem o qual este trabalho não teria sido possível. Agradecemos, ainda, ao Núcleo de Documentação Histórica da Universidade Federal de Pelotas (NDH-UFPel) pelo apoio institucional e pela disponibilização de conteúdos para o desenvolvimento desta pesquisa.

Referências

- Akgun, S. and Greenhow, C. (2022). Artificial intelligence in education: Addressing ethical challenges in k-12 settings. *AI and Ethics*, 2(3):431–440.
- Alario-Hoyos, C., Kemcha, R., Kloos, C. D., Callejo, P., Estévez-Ayres, I., Santín-Cristóbal, D., Cruz-Argudo, F., and López-Sánchez, J. L. (2024). Tailoring your code companion: Leveraging llms and rag to develop a chatbot to support students in a programming course. In *2024 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*, pages 1–8. IEEE.
- Anderson, J. R., Boyle, C. F., and Reiser, B. J. (1985). Intelligent tutoring systems. *Science*, 228(4698):456–462.
- Anthropic (2026). Claude 4.7 opus. Software / LLM. Versão 4.7.
- Bhatia, G., Peyrard, M., and Zhao, W. (2025). Date fragments: A hidden bottleneck of tokenization for temporal reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3201–3219.
- Chroma (2022). Chroma: the ai-native open-source embedding database. <https://www.trychroma.com>. Acesso em: Março de 2026.
- da Costa, V. K., Zen, E., da Silva, A. B. A., and Tavares, T. A. (2023). Experiências com uso de mesa tangível em espaços de educação não formal. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 983–994. SBC.
- Danuart, L., Mawardi, V. C., and Lee, V. (2024). Retrieval-augmented generation (rag) large language model for educational chatbot. In *2024 Ninth International Conference on Informatics and Computing (ICIC)*, pages 1–6. IEEE.
- de Sousa, R. R. A., Santos, A., and Diogo, A. Q. (2023). A preservação da memória histórica na era digital: Desafios, oportunidades e tecnologias interativas no ensino e na pesquisa. *Revista Inter-Ação*.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Halbwachs, M. (1997). *La mémoire collective*. Albin Michel.
- Idris, M. Z., Mustaffa, N. B., and Yusoff, S. O. S. (2016). Preservation of intangible cultural heritage using advance digital technology: Issues and challenges. *Harmonia: Journal of Arts Research and Education*, 16(1):1–13.
- Isotani, S., Bittencourt, I. I., Challco, G. C., Dermeval, D., and Mello, R. F. (2023). Aied unplugged: Leapfrogging the digital divide to reach the underserved. In *International Conference on Artificial Intelligence in Education*, pages 772–779. Springer.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

- Liu, Z., Agrawal, P., Singhal, S., Madaan, V., Kumar, M., and Verma, P. K. (2025). Lpi-tutor: an llm based personalized intelligent tutoring system using rag and prompt engineering. *PeerJ Computer Science*, 11:e2991.
- Mello, R. F., Anthony, L., Lobo, J., Ribeiro, F. G. C., Xavier, C., da Costa, N. T., Gasevic, D., and Rodrigues, L. (2025). Empowering equitable learning with llms: enhancing writing skills in low-resource contexts. In *European Conference on Technology Enhanced Learning*, pages 183–197. Springer.
- Microsoft (2024). Markitdown: Python tool for converting various files to markdown. <https://github.com/microsoft/markitdown>.
- Nora, P. and Erll, A. (1997). *Les lieux de mémoire*, volume 3. Gallimard Paris.
- Ollama (2023). Ollama. <https://ollama.com>. Acesso em: Março de 2026.
- Streamlit (2019). Streamlit: A faster way to build and share data apps. <https://streamlit.io>. Acesso em: Março de 2026.
- Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., et al. (2025). Gemma 3 technical report. *arXiv*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P. S., and Wen, Q. (2026). Large language models for education: A survey and outlook. *IEEE Signal Processing Magazine*, 42(6):51–63.
- Yin, J., Goh, T.-T., and Hu, Y. (2024). Interactions with educational chatbots: the impact of induced emotions and students’ learning motivation. *International Journal of Educational Technology in Higher Education*, 21(1):47.
- Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., et al. (2025). Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.