Conectando Dados a Decisões: Uma Proposta de Interface com Modelos de Linguagem para Gestores Educacionais

Abílio Nogueira Barros[1], **Lhaíslla Cavalcanti** [1], **Rafael Ferreira Mello**[1,2]

[1]Departamento de Computação - Universidade Federal Rural de Pernambuco (UFRPE)

[2]Centro de Estudos e Sistemas Avançados do Recife (CESAR)

abilionbarros@gmail.com, lhaislla.cavalcanti,rafael.mello@ufrpe.br

Resumo. *O acesso a dados públicos educacionais por gestores ainda enfrenta barreiras técnicas e operacionais, dificultando a incorporação de evidências na formulação de políticas e estratégias. A sobrecarga de atribuições administrativas, aliada à carência de pessoal técnico qualificado para manipulação de bases de dados, limita a capacidade dos gestores de obter respostas rápidas para perguntas simples do cotidiano escolar. Para enfrentar esse desafio, este trabalho propõe a criação de uma ferramenta baseada em Recuperação Aumentada com Geração (RAG), que permite a interação em linguagem natural com bases de dados educacionais, eliminando a dependência de painéis e dashboards complexos. Utilizando o framework Vanna.ai, foi desenvolvido um agente conversacional capaz de converter perguntas feitas em linguagem natural em consultas SQL executadas diretamente sobre bases públicas educacionais estruturadas. Essa abordagem promove maior confiabilidade nas respostas geradas, ao priorizar a extração direta da informação via consulta aos dados, sem interpretações sobre contexto textual.*

1. Introdução

Quando falamos em educação, especialmente na educação básica, etapa em que o ser humano passa quase as duas primeiras décadas de sua vida, diversos desafios se impõem sob a ótica de quem realiza a gestão de um sistema tão complexo. Problemas como evasão escolar [Neri et al. 2015] e retenção [Rebelo 2009], já amplamente conhecidos, tornaram-se ainda mais intensos nos últimos anos [Neri and Osorio 2021]. Além disso, a crise sanitária global de 2019 trouxe novos obstáculos, como a necessidade de adaptação dos professores ao ensino em ambientes virtuais [Sallaberry et al. 2020] e o desafio de manter a motivação de alunos e docentes durante o período pandêmico [da Costa et al. 2021].

Com o retorno às atividades presenciais, os educadores passaram a observar dificuldades significativas no convívio escolar, no cumprimento dos ritos acadêmicos e na socialização entre os estudantes [dos Santos Amaral and Martins 2023]. Tais desafios se mostraram especialmente graves nas etapas pré-escolares e de inserção ao ambiente educacional, afetando com maior intensidade as camadas socialmente mais vulneráveis, para as quais a escola, muitas vezes, representa inclusive a principal fonte de alimentação [Freitas 2023].

Diante da junção de problemas antigos e emergentes que impactam o cotidiano escolar, a tomada de decisões no contexto educacional torna-se uma tarefa ainda mais

DOI: 10.5753/wapla.2025.15939

sensível. Isso exige embasamento sólido, sob o risco de gerar desperdício de recursos ou de negligenciar públicos menos assistidos. No entanto, muitos gestores e tomadores de decisão enfrentam limitações no acesso a dados públicos educacionais de forma estruturada, clara e acessível. A ausência de pessoal técnico dedicado exclusivamente à manipulação e interpretação desses dados, associada à sobrecarga das funções administrativas, compromete a capacidade de incorporar evidências quantitativas de maneira sistemática no processo decisório.

Nesse contexto, torna-se necessário o desenvolvimento de mecanismos que ampliem a acessibilidade aos dados educacionais, sem depender de interfaces complexas ou *dashboards* analíticos. Apesar da popularização de painéis interativos, muitos deles exigem conhecimento técnico para interpretação e manutenção contínua. Além disso, apresentam limitações na capacidade de responder com precisão a perguntas pontuais ou exploratórias que surgem no cotidiano da gestão, como *"quantas escolas de tempo integral temos em determinada região?"* ou *"qual o número de matrículas por dependência administrativa no último censo escolar?"*. Para superar tais limitações, propõe-se uma abordagem mais direta e responsiva, fundamentada no uso de agentes conversacionais, que permita o acesso natural à informação, sem a mediação de interfaces visuais tradicionais.

O presente trabalho busca demonstrar a criação do conjunto de dados inicial e a ferramenta de consulta aos dados com base no *framework Vanna.AI*, definindo os passos necessários para replicabilidade dessa ferramenta e seu treinamento coerente.

2. Contextualização

Agentes de IA já vem sendo utilizados para auxílio de *feedback* em sala de aula, principalmente utilizando o *Generative Pre-trained Transformer (GPT)*, esses agentes buscam resolver problemas como personalização, motivação e escalabilidade na educação, adaptando conteúdo e *feedback* às necessidades individuais dos alunos [Diniz et al. 2022, Neto et al. 2023, Córdova-Esparza 2025]

A utilização de agentes conversacionais integrados a bases educacionais públicas emerge como uma alternativa viável e promissora em razão de sua capacidade de escalabilidade, disponibilidade contínua e adaptabilidade a diferentes perfis de usuários. Combinando processamento de linguagem natural e acesso direto às bases por meio de linguagens de consulta estruturadas, é possível criar um ambiente onde perguntas em linguagem natural são automaticamente traduzidas em consultas *Structured Query Language (SQL)* precisas. Essa abordagem viabiliza uma tomada de decisão orientada por dados mesmo por profissionais sem conhecimento técnico em análise de dados ou programação.

Agentes de IA já vêm sendo empregados para auxiliar no feedback em sala de aula, principalmente utilizando o GPT, esses agentes buscam resolver problemas como personalização, motivação e escalabilidade na educação, adaptando conteúdo e *feedback* às necessidades individuais dos alunos [Diniz et al. 2022, Neto et al. 2023, Córdova-Esparza 2025]

O *Retrieval-Augmented Generation (RAG)* é uma solução que supera a principal barreira para a adoção de *chatbots* baseados em LLMs na educação: as alucinações. Ele foi introduzido como uma abordagem capaz de lidar com a dificuldade de atualizar ou

estender o conhecimento dos LLMs e sua propensão a gerar texto sem sentido ou infiel à fonte. O RAG aprimora o desempenho dos LLMs integrando técnicas de recuperação de informações. Em vez de otimizar o treinamento do modelo, ele usa a entrada do usuário para recuperar documentos de texto relevantes de bases de dados externas, que são utilizadas como contexto adicional para gerar a resposta [Gupta et al. 2024, Gao et al. 2023].

3. Metodologia

A metodologia adotada neste estudo foi a *Design Research*, com o objetivo principal de desenvolver uma ferramenta capaz de integrar uma bases de dados educacionais a uma interface conversacional, permitindo assim a validação do processo de interação entre o usuário e os dados. Essa abordagem foi escolhida por sua capacidade de aliar a construção prática de soluções com a investigação científica voltada à resolução de problemas reais.

O processo metodológico foi dividido em duas etapas principais como pode ser visto na Figura 1. A primeira etapa consistiu na construção de uma base de dados estruturada, capaz de agregar inicialmente indicadores educacionais do ensino básico e informações sobre as unidades escolares, com foco em atender às necessidades de gestores escolares. O intuito foi reunir, organizar e tornar acessíveis dados relevantes que subsidiem a tomada de decisão no contexto educacional.

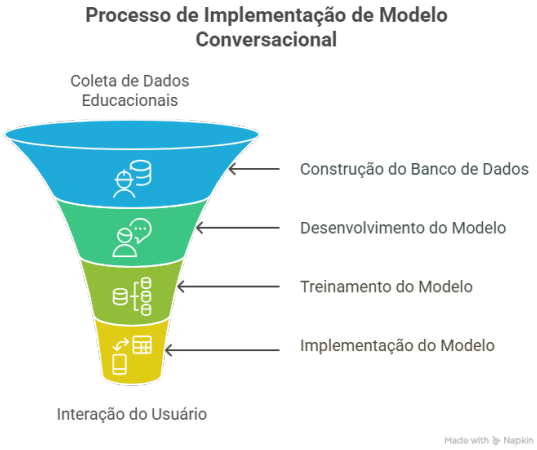

Figura 1: Fluxo geral da solução.

A segunda etapa envolveu o desenvolvimento da *pipeline* de criação, treinamento e implementação do modelo conversacional. Esse modelo foi integrado à interface da plataforma distribuidora, possibilitando que os usuários finais, neste caso, os gestores escolares, possam interagir diretamente com os dados por meio de uma interface intuitiva, baseada em linguagem natural.

3.1. Desenvolvimento da base de dados

Sendo o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) a principal fonte de divulgação de dados educacionais aberto do Brasil, foi realizada uma análise no formato de divulgação dos dados, onde o Censo da Educação Básica traz informações condensadas em nível escolar sobre as instituições brasileiras. Partindo desse

ponto, foram escolhidos também três indicadores, Taxas de aproveitamento (Aprovação, Reprovação e Abandono),Taxa de distorção idade-série (TDI) e o indicador de complexidade da gestão escolar (ICG). Esses indicadores que são em nível escolar e podem ser comparados durante o período analisado.

Os dados, tanto em nível de microdados, que são o Censo da educação básica, quanto os indicadores selecionados, foram extraídos do site do INEP, nas seções de Microdados[1] e Indicadores educacionais da educação Básica[2].

Foram selecionados três indicadores para esta primeira versão da aplicação. A escolha foi baseada em sua relevância para a análise da trajetória acadêmica dos alunos e na identificação de dificuldades no processo escolar, possibilitando uma mensuração eficaz tanto para a gestão escolar quanto para as secretarias distritais ou municipais de educação.

As taxas de rendimento representam indicadores educacionais que consolidam os percentuais de aprovação, reprovação e abandono escolar por instituição de ensino. Esses dados, disponibilizados anualmente por etapa e ano escolar a partir do 1º ano do ensino fundamental, oferecem um panorama da trajetória dos estudantes ao longo do ano letivo. A taxa de distorção idade-série (TDI) indica a proporção de alunos matriculados em séries inadequadas à sua idade, com defasagem superior a dois anos. Esse indicador é fundamental para compreender os desafios do fluxo escolar, pois reflete problemas como evasão, retenção e os custos associados à permanência prolongada dos alunos nas etapas de ensino.

O Indicador de Complexidade da Gestão Escolar (ICG) classifica as escolas em uma escala de 1 a 6, de acordo com o grau de complexidade da gestão institucional. Esse grau é determinado a partir de variáveis como o porte da escola, o número de turnos de funcionamento, a diversidade de etapas e modalidades de ensino oferecidas, além da presença de turmas com mediação pedagógica semipresencial ou à distância. Níveis mais altos indicam maior complexidade e, consequentemente, maiores desafios administrativos. Essa classificação auxilia na compreensão das diferentes realidades enfrentadas pelas escolas na gestão de seus recursos e na organização do ensino.

A base do Censo Escolar foi selecionada para esta aplicação devido à sua abrangência e relevância para a análise educacional, porém exigiu adaptações metodológicas.O processamento dos indicadores e microdados foram desenvolvidos com base no processo e ferramental adaptados dos trabalhos de [Barros et al. 2022] e [Barros et al. 2023]. Foram necessários ajustes quanto ao tratamento dos dados, especialmente no que diz respeito a variáveis qualitativas que passaram a aparecer com valores zero, mesmo em anos nos quais ainda não eram coletadas. Para garantir a integridade das análises, tornou-se essencial identificar não apenas a presença das colunas nos arquivos, mas também a efetiva coleta das variáveis em cada ano, evitando interpretações equivocadas sobre a ausência ou existência de determinadas características nas escolas.

Para a disponibilização dos dados, o *DuckDB* foi escolhido por seu algoritmo de compressão, que permitiu armazenar a base de dados de 2018 a 2023. Outra etapa importante foi o desenvolvimento do dicionário de dados da base, incluindo as adaptações

[1]https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados
[2]https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais

necessárias durante a etapa de pré processamento. Para os arquivos de indicadores os quais não existiam descrição foram desenvolvidos para a etapa seguinte.Os dados e metadados podem ser acessados pelo ZENODO.

3.2. Implementação do Modelo conversacional

Para este experimento, foi utilizado o *framework* Vanna.ai, um pacote *python open-source* que facilita a combinação entre banco de dados , grande modelos de linguagem. A combinação dessas ferramentas juntas a uma interface de consulta para o usuário provê um agente conversacional baseado em RAG *(Retrieval-Augmented Generation)*, que permite a implementação simplificada de interfaces conversacionais conectadas a bases relacionais.

O *Vanna* atua como uma ponte entre perguntas formuladas em linguagem natural e a geração de *queries SQL* que consultam diretamente os dados subjacentes. Sua arquitetura modular e o suporte a modelos abertos de linguagem o tornam uma escolha adequada para ambientes educacionais públicos, onde a infraestrutura pode ser limitada.

Diferentemente de abordagens que utilizam documentos ou *embeddings* de contexto, o *Vanna.ai* realiza a recuperação da informação diretamente por meio de *queries SQL* sobre os dados brutos. Isso eleva a confiabilidade das respostas, pois evita interpretações ambíguas ou construções narrativas sobre documentos textuais, priorizando a consulta exata sobre dados estruturados. O processo de extração de respostas com o *Vanna.ai* pode ser avaliado na figura 2.

Figura 2: Fluxo de funcionamento do Vanna.ai.

A *pipeline* desenvolvida nesta implementação integra uma base de dados desenvolvida, garantindo sua integridade e utilizando os metadados produzidos também como primeira fonte de conhecimento para o modelo. O agente *Vanna* foi configurado com um modelo de linguagem ajustado para a tarefa, utilizando, em um primeiro momento, um modelo gratuito com limite de acesso diário. Todo o processo é orquestrado em ambiente controlado. Isso permite tanto a realização de testes manuais na própria interface do

framework, quanto a integração com outras ferramentas *Python*, como o *Streamlit*, o que viabiliza uma futura implementação em sistemas públicos.

4. Resultados

Os resultados demonstram o desenvolvimento e a funcionalidade do agente conversacional proposto para a gestão educacional. A análise da ferramenta se divide em duas partes principais: a primeira aborda a capacidade da interface de traduzir a linguagem natural para consultas *SQL*, enquanto a segunda foca na apresentação dos dados e nas possibilidades de análise oferecidas.

A Figura 3 ilustra um exemplo de interação entre o usuário e o agente conversacional, demonstrando o fluxo desde a pergunta em linguagem natural até a obtenção da resposta. O usuário inicia a interação com a pergunta: "Qual a média do TDI das escolas nos finais do ensino fundamental?" e o agente, por sua vez, utiliza a *pipeline* de processamento do *Vanna.ai*, interpreta a pergunta, identifica os termos relevantes e os converte em uma consulta *SQL*. A consulta é então executada na base de dados, retornando o valor numérico que corresponde à média da taxa de distorção idade-série para as escolas na faixa temporal analisada.

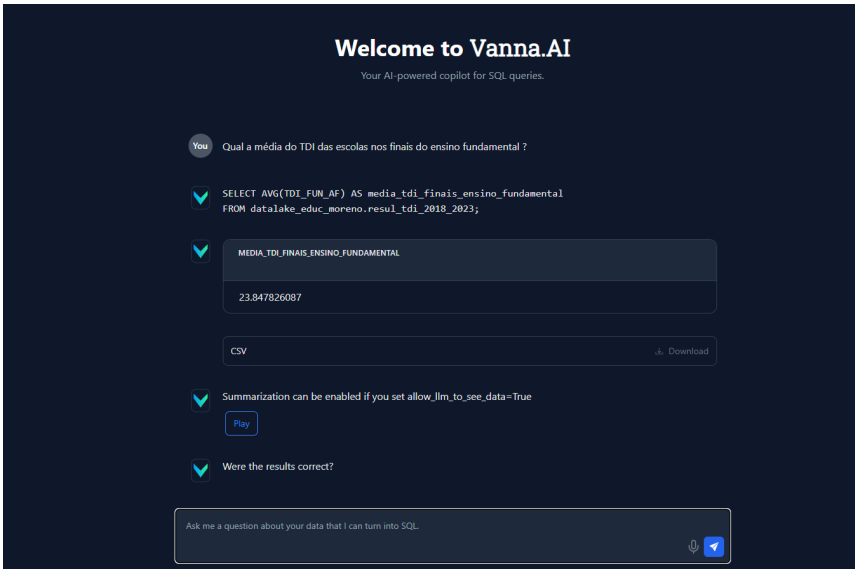

Figura 3: Execução do agente conversacional.

Este processo de obtenção de resposta demonstra a capacidade do agente em abstrair a complexidade técnica, traduzindo a intenção do usuário em uma consulta estruturada. Este processo garante que a resposta fornecida seja diretamente extraída da fonte de dados, eliminando ambiguidades e a necessidade de conhecimento em linguagem de banco de dados por parte do gestor.

A Figura 4 retrata outra funcionalidade, onde o agente não só responde à pergunta de forma simples, mas também consegue realizar análises, retornando dados estruturados em formato de tabela que podem ser exportados para um arquivo, como *CSV*, o que facilita o uso dos dados em outras ferramentas, como planilhas eletrônicas, para análises detalhadas.

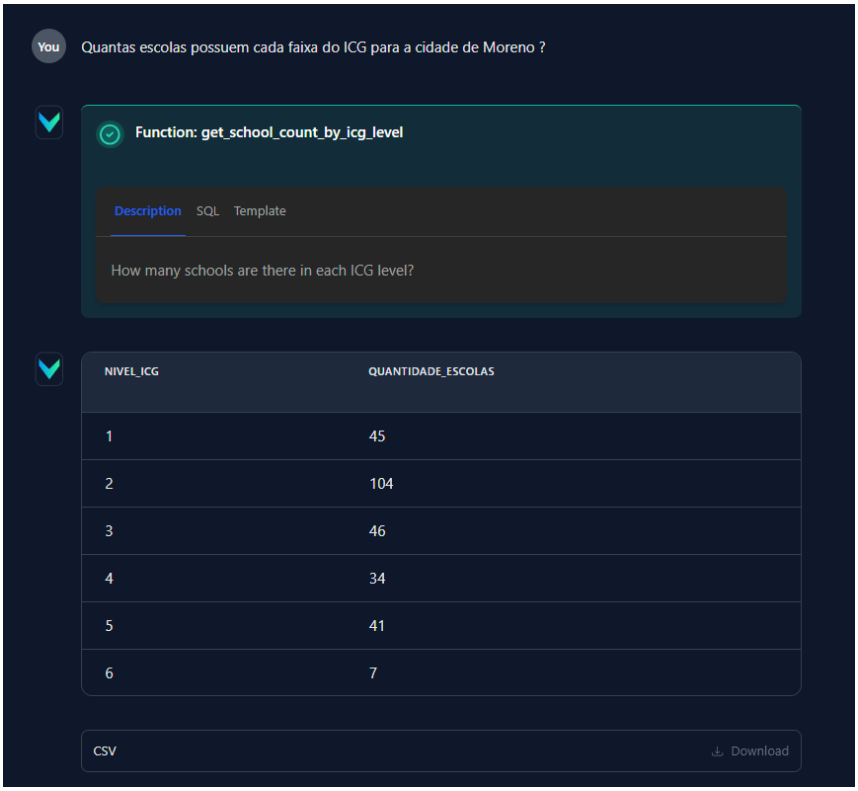

Figura 4: Exemplo de pergunta e formas de saída da resposta.

A aplicação deste tipo de ferramenta no contexto educacional demonstra a democratização do acesso à informação, à capacidade de fornecer respostas precisas e a possibilidade de gerar relatórios. Esses benefícios permitem que gestores tomem decisões rápidas e embasadas em evidências sem a necessidade de habilidades técnicas avançadas.

Um dos principais pontos fortes desta abordagem é sua capacidade de adaptação a outras ferramentas baseadas na linguagem *Python*, uma linguagem de código aberto amplamente adotada globalmente, conforme demonstrado pelo índice TIOBE. Essa flexibilidade permite a integração com diversos *frameworks* para desenvolvimento de interfaces, como *Flask*, *Django* e *Streamlit*, facilitando sua incorporação a sistemas já existentes ou a soluções personalizadas.

Outro aspecto relevante é a possibilidade de integração com bases de dados locais. A estrutura proposta permite adicionar dados de forma relacional às informações utilizadas no estudo, incluindo dados produzidos por municípios, redes de ensino ou escolas específicas. Desde que esses dados sejam organizados conforme os princípios básicos de estruturação e disponibilização, como elencados na metodologia desse trabalho.

Sua utilização continua faz com que a ferramenta possar ter um melhor desempenho visto que é possível salvar perguntas frequentes criando funções. Mesmo esse não sendo o objetivo principal da ferramenta, trás mais uma possibilidade que é a resposta rápida a perguntas frequentes. Essas perguntas novamente são usadas para o treinamento do modelo e podem ser exportadas para implementação em outras instituições de uma mesma rede escolar, por exemplo, fazendo com que um primeiro conjunto de gestores escolares sirvam de base para implementação nas demais escolas.

No entanto, faz-se necessário reconhecer limitações, como a dependência da qualidade e atualização da base de dados subjacente, uma vez que a ferramenta só será eficaz se os dados sejam eles internos ou externos estiverem corretos e bem estruturados. Além disso, a interpretação de perguntas excessivamente complexas ou ambíguas ainda pode representar um desafio. Apesar dessas ressalvas, a abordagem apresentada valida um caminho para integrar a análise de dados no cotidiano da gestão educacional de forma prática.

Outro ponto de atenção diz respeito ao modelo de linguagem escolhido. No caso de modelos pagos, é essencial avaliar os custos de uso em função da aceitação da ferramenta pelos gestores. A implementação sem planejamento adequado pode inviabilizar essa e outras estratégias futuras de adoção de tecnologias para decisões baseadas em dados. A escolha do modelo mais apropriado deve considerar tanto a realidade em que será aplicado quanto os objetivos pretendidos. Isso evita desperdício de recursos e garante que o modelo atenda com precisão e agilidade às demandas especialmente em contextos de perguntas complexas, reduzindo o risco de descontinuidade da ferramenta.

5. Considerações Finais

Este trabalho desenvolveu e aplicou uma ferramenta baseada em RAG para facilitar o acesso de gestores educacionais a dados públicos estruturados. A solução traduz perguntas em linguagem natural para consultas SQL, permitindo respostas rápidas e confiáveis, sem exigir conhecimentos técnicos em manipulação de bases de dados.

A proposta se destacou pela flexibilidade de integração com outras ferramentas e bases locais, além de permitir personalizações conforme as necessidades específicas de cada rede de ensino. A possibilidade de reaproveitamento de perguntas frequentes e a adaptabilidade à linguagem *Python* fortalecem sua escalabilidade e aplicabilidade em diferentes contextos educacionais.

5.1. Trabalhos futuros

A validação com gestores é essencial em iniciativas que visam consolidar uma cultura baseada em dados. A continuidade deste trabalho se dará em duas etapas principais. Primeiramente, aplicar de questionários após a apresentação da ferramenta, visando identificar as principais perguntas dos gestores, avaliar a utilidade dos *dashboards* existentes e refinar o público-alvo da solução. Isso também permitirá aprimorar o modelo para oferecer respostas confiáveis e com redução de latência já no primeiro uso.

A segunda etapa envolverá testes com usuários para avaliar a integração do chat em uma nova interface ou em interfaces já existentes. Buscando aumentar a simplicidade no uso da ferramenta e com isso uma maior aceitação dos usuários, fortalecendo assim o conceito de utilização de ferramentas para suporte à tomada de decisão.

Outros aspectos de melhoria, como a avaliação dos modelos utilizados e o monitoramento de custos operacionais, serão tratados a partir dos primeiros ciclos de uso, respeitando as especificidades de cada rede. Além disso, a viabilidade de treinamentos direcionados à reformulação das perguntas buscando maior objetividade poderá permitir o uso de modelos mais simples, reduzindo custos e reforçando a adoção sustentável da ferramenta.

Referências

[Barros et al. 2022] Barros, A. N., Alencar, A., Nascimento, A., de Albuquerque, A. F., and Mello, R. F. (2022). Elaboração do conjunto de dados agregados do censo da educação básica. In *Anais do IV Dataset Showcase Workshop*, pages 35–45. SBC.

[Barros et al. 2023] Barros, A. N., de Alencar, A. L., and Mello, R. F. (2023). Criação de um panorama da taxa de distorção idade-série com base nos dados abertos do inep. In *Congresso sobre Tecnologias na Educação (Ctrl+ e)*, pages 457–462. SBC.

[Córdova-Esparza 2025] Córdova-Esparza, D.-M. (2025). Ai-powered educational agents: Opportunities, innovations, and ethical challenges. *Information*, 16(6):469.

[da Costa et al. 2021] da Costa, H. C. O., de Carvalho, A. d. S. M., dos Santos, T. S., and Pereira, P. C. (2021). Motivação para ensinar e aprender em tempo de pandemia. *Research, Society and Development*, 10(16):e558101624122–e558101624122.

[Diniz et al. 2022] Diniz, G., Alencar, E., and Nunes, I. (2022). Monitores automáticos - uso de learning analytics para acompanhamento de atividades. In *Anais do I Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*, pages 79–87, Porto Alegre, RS, Brasil. SBC.

[dos Santos Amaral and Martins 2023] dos Santos Amaral, M. and Martins, V. L. (2023). Desafios da educação infantil pós-pandemia: Breve análise de algumas dificuldades enfrentadas na retomada presencial. *A educação na contemporaneidade: desafios pedagógicos e tecnológicos–Volume 2*, page 38.

[Freitas 2023] Freitas, P. R. D. M. (2023). A educação pós-pandemia covid 19–a retomada das aulas presenciais ganhos e perdas na aprendizagem. *Gestão & Educação*, page 88.

[Gao et al. 2023] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., and Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

[Gupta et al. 2024] Gupta, S., Ranjan, R., and Singh, S. N. (2024). A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions. *arXiv preprint arXiv:2410.12837*.

[Neri et al. 2015] Neri, M. et al. (2015). Motivos da evasão escolar. *Centro de Políticas Sociais*.

[Neri and Osorio 2021] Neri, M. and Osorio, M. C. (2021). Evasão escolar e jornada remota na pandemia. *Revista NECAT-Revista do Núcleo de Estudos de Economia Catarinense*, 10(19):28–55.

[Neto et al. 2023] Neto, R., Alves, G., and Mello, R. (2023). Aplicando chatgpt para recomendação de tags para auxiliar professores na correção de atividades abertas. In *Anais do II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*, pages 51–60, Porto Alegre, RS, Brasil. SBC.

[Rebelo 2009] Rebelo, J. A. (2009). Efeitos da retenção escolar, segundo os estudos científicos, e orientações para uma intervenção eficaz: Uma revisão. *Revista portuguesa de pedagogia*, pages 27–52.

[Sallaberry et al. 2020] Sallaberry, J. D., dos Santos, E. A., Bagatoli, G. C., Lima, P. C. M., and Bittencourt, B. R. (2020). Desafios docentes em tempos de isolamento social:

estudo com professores do curso de ciências contábeis. *Revista Docência do Ensino Superior*, 10:1–22.