

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Otimização do processo de ensino no Brasil através do uso de IA

Daniel da Silva Amaral

Monografia - MBA em Inteligência Artificial e Big Data

**UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO**

Daniel da Silva Amaral

**Otimização do processo de ensino no Brasil através do uso
de IA**

São Carlos

2024

Daniel da Silva Amaral

Otimização do processo de ensino no Brasil através do uso de IA

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. André Ricardo Backes

Versão original

São Carlos

2024

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTA TRABALHO,
POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO PARA FINS DE ESTUDO E
PESQUISA, DESDE QUE CITADA A FONTE.

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados
fornecidos pelo(a) autor(a)

Silva Amaral, Daniel da

Otimização do processo de ensino no Brasil através do uso
de IA / Daniel da Silva Amaral ; orientador André Ricardo
Backes. – São Carlos, 2024.

48 p. : il. (algumas color.) ; 30 cm.

Monografia (MBA em Inteligência Artificial e Big Data) –
Instituto de Ciências Matemáticas e de Computação, Universi-
dade de São Paulo, 2024.

1. LaTeX. 2. abnTeX. 3. Classe USPSC. 4. Editoração de
texto. 5. Normalização da documentação. 6. Tese. 7. Disserta-
ção. 8. Documentos (elaboração). 9. Documentos eletrônicos.
I. Ricardo Backes, André, orient. II. Título.

RESUMO

AMARAL, D.S. **Otimização do processo de ensino no Brasil através do uso de IA.** 2024. 48 p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

A educação é um dos pilares fundamentais do desenvolvimento de qualquer sociedade. No Brasil busca-se superar obstáculos históricos relacionados à qualidade e à equidade no ensino. Neste contexto, o uso de tecnologias inovadoras, como a Inteligência Artificial, surge como um agente de transformação, especialmente em áreas como a personalização do aprendizado e no suporte ao corpo acadêmico.

Com a ascensão dos *Large Language Models*, a Inteligência Artificial Generativa se tornou uma ferramenta com alto potencial para a criação de sistemas que tem como intuito otimizar o processo de ensino.

Este trabalho tem como objetivo apresentar uma solução inovadora para a otimização do processo de ensino no Brasil através do desenvolvimento de um assistente inteligente para suporte ao ensino.

A proposta visa, de forma prática, abordar as seguintes questões: Como a IA generativa pode contribuir para a criação de um assistente virtual eficiente para o aprendizado? Quais são os principais desafios e oportunidades na implementação de um assistente virtual em um contexto educacional? Que tipo de impacto essa solução pode gerar para a qualidade do ensino e para a experiência do estudante?

O desenvolvimento da solução explorado neste trabalho envolve a seleção de um LLM apropriado, a criação de um sistema de processamento de dados e a implementação de uma interface para o usuário. O trabalho também apresenta os resultados de testes realizados com a solução desenvolvida a título de MVP, avaliando sua efetividade e identificando as áreas de melhoria.

A partir da análise dos resultados e da discussão dos desafios e oportunidades encontrados, este trabalho propõe um estudo aprofundado sobre o papel da IA na educação, buscando contribuir para a construção de um futuro mais promissor para o ensino no Brasil.

ABSTRACT

AMARAL, D.S. **Optimization of the teaching process in Brazil through the use of AI.** 2024. 48 p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

Education is one of the fundamental pillars of the development of any society. In Brazil, the aim is to overcome historical obstacles related to quality and equity in education. In this context, the use of innovative technologies, such as Artificial Intelligence, emerges as an agent of transformation, especially in areas such as personalization of learning and support for academic staff.

With the rise of Large Language Models, Generative Artificial Intelligence has become a powerful tool for creating systems that aim to optimize the teaching process.

This work aims to present an innovative solution for optimizing the teaching process in Brazil through the development of an intelligent assistant to support teaching.

The proposal aims to address the following questions in a practical way: How can generative AI contribute to the creation of an efficient virtual assistant for learning?; What are the main challenges and opportunities in implementing a virtual assistant in an educational context?; What kind of impact can this solution generate for the quality of teaching and the student experience?

The development of the solution explored in this work involves the selection of an appropriate LLM, the creation of a data processing system and the implementation of a user interface. The work also presents the results of tests carried out with the solution developed as an MVP, evaluating its effectiveness and identifying areas for improvement.

Based on the analysis of the results and the discussion of the challenges and opportunities encountered, this work proposes an in-depth study on the role of AI in education, seeking to contribute to the construction of a more promising future for education in Brazil.

Keywords: Artificial Intelligence, Education, Educational Assistant, LLM, Language Models, Personalization of Learning.

LISTA DE FIGURAS

Figura 1 – Desempenho estudantes brasileiros nos 20 anos do índice PISA. Fonte: (OECD, 2024, p4)	12
Figura 2 – Diagrama simplificado Retrieval-Augmented Generation	15
Figura 3 – Diagrama dos macro componentes da solução	16
Figura 4 – Framework EduChat. Fonte: (Dan <i>et al.</i> , 2023, p4)	20
Figura 5 – Fluxo atividades concepção trabalho.	21
Figura 6 – Comparação de Modelos - quesito qualidade. Fonte: (Analysis, 2024)	24
Figura 7 – General Ability Chatbot Arena. Fonte: (Analysis, 2024)	25
Figura 8 – Reasoning e Knowledge MMLU. Fonte: (Analysis, 2024)	26
Figura 9 – Coding HumanEval. Fonte: (Analysis, 2024)	27
Figura 10 – Comparação de Modelos - quesito velocidade. Fonte: (Analysis, 2024)	28
Figura 11 – Comparação de Modelos - quesito preço. Fonte: (Analysis, 2024)	28
Figura 12 – Benchmark consolidado baseado em (Analysis, 2024)	29
Figura 13 – Arquitetura de Dados.	30
Figura 14 – Arquitetura Aplicação	31
Figura 15 – Arquitetura MVP	32
Figura 16 – Resposta prompt interação inicial	35
Figura 17 – Resposta pergunta sobre conteúdo	36
Figura 18 – Consulta referências no material didático	37
Figura 19 – Dúvida fórmula	37
Figura 20 – Resposta dúvida fórmula	38
Figura 21 – Dúvida via áudio	39
Figura 22 – Pesquisa quantitativa parte 1	40
Figura 23 – Pesquisa quantitativa parte 2	41
Figura 24 – Pesquisa quantitativa parte 3	42
Figura 25 – Pesquisa quantitativa parte 4	43
Figura 26 – Pesquisa quantitativa parte 5	43
Figura 27 – Público	44

SUMÁRIO

1	INTRODUÇÃO	11
1.1	Motivação	12
2	FUNDAMENTAÇÃO TEÓRICA	13
2.1	Breve histórico de experimentos conversacionais entre humanos e computadores	13
2.2	Large Language Models (LLMs)	14
2.3	Lidando com conhecimento estático de LLMs e mitigando alucinações	14
2.3.1	Retrieval-Augmented Generation (RAG)	15
2.3.2	Chain-of-Thought (CoT)	15
2.4	Componentes complementares	16
3	ESTADO DA ARTE	19
4	METODOLOGIA	21
5	DESENVOLVIMENTO	23
5.1	Escolha do LLM	23
5.2	Arquitetura de Dados e Aplicação	30
5.3	Recorte do escopo do projeto para fins de desenvolvimento do MVP	32
6	COLETA DE RESULTADOS OBTIDOS	33
7	CONCLUSÃO	45
	REFERÊNCIAS	47

1 INTRODUÇÃO

A educação é um dos pilares fundamentais no desenvolvimento de qualquer país. A médio e longo prazo, a adoção de uma estratégia de ensino eficaz pode impactar drasticamente na economia, saúde e outras áreas.

Medir a eficácia do sistema de ensino é uma tarefa complexa, porém existem diversos índices que podem ajudar nesta avaliação. O Índice de Desenvolvimento Humano, ou IDH, é o principal índice global para classificação do nível de desenvolvimento de um país. Este é composto por 3 indicadores: Renda, Saúde e Educação. Dentro do critério de educação, são avaliados os graus de instrução, taxas de alfabetização, taxas de escolaridade, evasão escolar, taxas de repetência entre outras variáveis. Segundo a última publicação, realizada em 2022, o Brasil ficou na posição 87 de 191 ((PNUD), 2022).

Um outro índice global, mas agora restrito ao quesito educação, é o gerado a partir do Programa de Avaliação de Estudantes, ou PISA. O mesmo teve sua publicação mais recente realizada em 2022 e a posição do Brasil se manteve estável em relação à avaliação anterior, ocorrida em 2018. De acordo com os resultados da prova aplicada globalmente em estudantes de 15 anos de idade, cerca de 73% dos estudantes brasileiros não conseguem converter moedas ou comparar distâncias de forma adequada (Tenente, 2023).

A fim de contemplar a perspectiva nacional da qualidade de ensino no país, pode-se observar o Índice de Desenvolvimento da Educação Básica, o IDEB, cuja última publicação se deu em 2021. Nessa última publicação os resultados atribuídos ao Ensino Fundamental - Anos Iniciais, Ensino Fundamental - Anos Finais e Ensino Médio foram respectivamente de 5.8/10, 5.1/10 e 4.2/10 (Teixeira, 2021).

Os números apresentados revelam apenas sintomas da ineficiência do ensino no país, mas vale observar que muitas causas podem colaborar para o cenário exposto. Entre elas, podemos citar:

- Desigualdade social
- Infraestrutura precária das redes de ensino
- Escassez de professores qualificados
- Metodologias de ensino ultrapassadas

A proposta desse projeto visa o desenvolvimento de uma solução de inteligência artificial que tem por meta atacar os dois últimos itens da lista acima através das seguintes estratégias:

- Colaborar com a disseminação do conhecimento de forma massificada e acessível a todos os estudantes através do desenvolvimento de Modelos de Aprendizado Profundo e Processamento de Linguagem natural treinados com material alinhado a base curricular nacional.
- Hiper personalizar a abordagem de ensino de acordo com as habilidades ou mesmo inabilidades de cada aluno.
- Usar técnicas de Reforço de Aprendizado a fim de otimizar constantemente o modelo através de instrumentos de avaliação particulares de cada instituição e também instrumentos públicos como os mencionados na seção de contextualização.

1.1 Motivação

De acordo com os números apresentados, a evolução dos índices da qualidade e eficácia do ensino do Brasil representam um grande desafio para a nação conforme figura 1 extraída do relatório PISA 2022

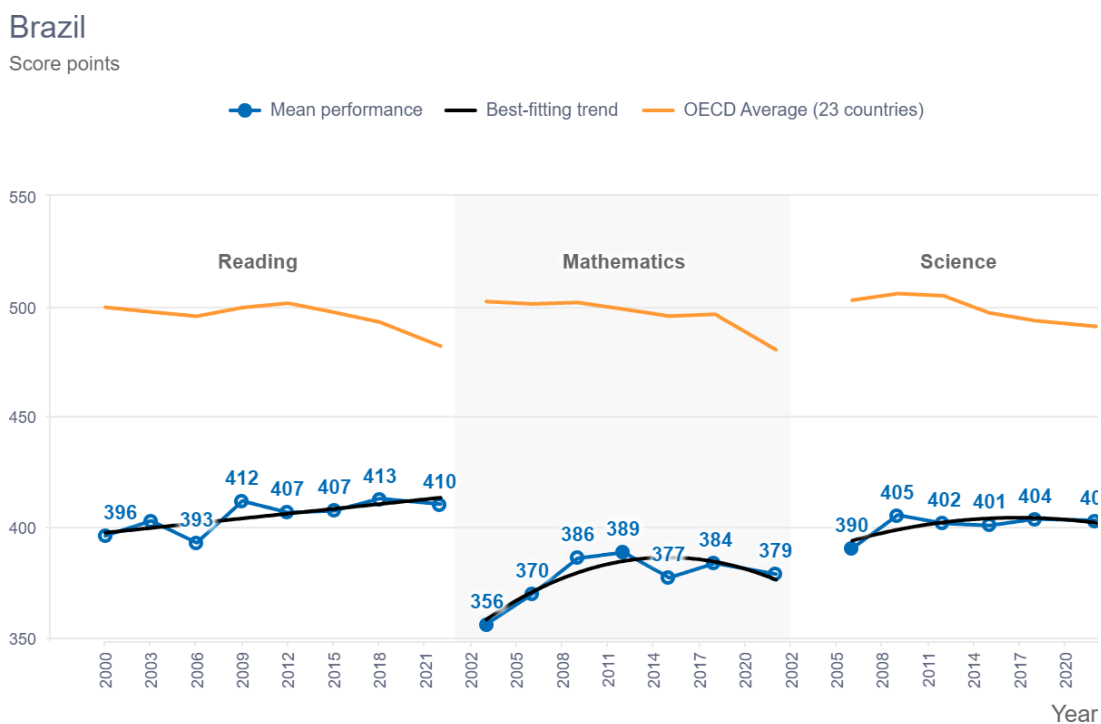


Figura 1 – Desempenho estudantes brasileiros nos 20 anos do índice PISA. Fonte: (OECD, 2024, p4)

A hipótese em avaliação é de que o rendimento dos estudantes pode ser otimizado largamente através do uso de um assistente educacional digital durante a vida letiva. Esta solução tem como objetivo ajudar cada aluno de forma personalizada, de acordo com suas peculiaridades e histórico estudantil.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Breve histórico de experimentos conversacionais entre humanos e computadores

Soluções de software que fazem uso de processamento de linguagem natural para facilitar a interação entre humanos e máquinas não são nenhuma novidade. Já nos anos 60, (Weizenbaum, 1966) apresentou ao mundo Eliza, um dos primeiros *bots* criados com a intenção de simular experiências conversacionais.

Com o passar do tempo, o surgimento das técnicas de aprendizado profundo e redes neurais trouxeram grandes benefícios para o PLN (Processamento de Linguagem Natural), aprimorando significativamente a capacidade de processar e entender a linguagem natural (Amaratunga, 2023, p. 21).

Em 2017 pesquisadores do Google publicaram um artigo apresentando um novo padrão de arquitetura denominado *Transformers*, que viria a revolucionar o campo de Processamento de Linguagem Natural (Vaswani *et al.*, 2017).

Em 2018 foi introduzida pelo Google uma primeira família de modelos de linguagem baseada na arquitetura *Transformers* a *Bidirecional Encoder Representations from Transformers* ou *BERT* (Devlin *et al.*, 2019). Na sequência, em 2019 foi lançado o *Text-to-Text Transfer Transformer* ou *T5* (Raffel *et al.*, 2023), seguido pelo lançamento em 2020 da primeira versão do modelo *LaMDA* (*Language Model for Dialogue Applications*) (Google, 2021) batizado de *Meena* (Adiwardana *et al.*, 2020).

Alguns anos depois, essa revolução originada pela arquitetura *Transformers*, passou a ficar acessível ao grande público através do *Chat Generative Pre-trained Transformer* ou Chat GPT (OpenAI, 2022). Através do Chat GPT qualquer pessoa com uma conexão a internet passou a ter acesso a uma solução de inteligência artificial generativa de forma simplificada.

Outras soluções baseadas em *Large Language Models* foram lançadas pouco tempo depois do Chat GPT como, por exemplo, o Gemini (inicialmente chamado de Bard) do Google (Google, 2023) e o LLaMA da Meta (Meta, 2023)

É dentro deste contexto que concebemos esta proposta de solução baseada em inteligência artificial para suporte ao ensino no Brasil.

2.2 Large Language Models (LLMs)

Segundo (Amaratunga, 2023, p. 73), *Large Language Models* (LLMs) são modelos de inteligência artificial (IA) com capacidade de processar e interpretar a linguagem natural. Eles são compostos por um grande número de parâmetros e treinados em um conjunto de dados massivo. Essa característica os diferencia de outros modelos, permitindo que aprendam com mais complexidade e contexto.

A arquitetura *Transformer* é frequentemente utilizada em LLMs devido à sua eficiência no processamento de sequências de texto. Os parâmetros, nesse contexto, representam os “conhecimentos” que o modelo adquire durante o treinamento. Uma maior quantidade de parâmetros tende a ser uma métrica importante na medição da capacidade do modelo de aprender e se adaptar a diferentes tarefas. LLMs modernos como o Gemini ou GPT 4 podem ter centenas de bilhões de parâmetros.

2.3 Lidando com conhecimento estático de LLMs e mitigando alucinações

Conforme mencionado anteriormente, LLMs são modelos treinados em conjuntos massivos de dados. Não obstante, todo esse conhecimento está restrito a data na qual o *corpus* de dados usado para treino foi elaborado. Perguntas relacionadas a uma informação posterior a data do *corpus* utilizado, ou mesmo que não tenha sido consideradas na composição do mesmo, tendem a gerar um comportamento conhecido como alucinação.

Além disso, ainda que o LLM possua conhecimento para atender a uma solicitação, a maneira como esta é elaborada (*prompt*) pode influenciar drasticamente na resposta do mesmo.

Existe uma série de estratégias que podem ser adotadas para mitigar o risco de alucinação, fundamentar respostas (*grounding*) e otimizar solicitações para um *LLM*. *Retrieval-Augmented Generation (RAG)* e *Chain-of-Thought (CoT)* são dois exemplos dessas estratégias.

2.3.1 Retrieval-Augmented Generation (RAG)

Trata-se uma técnica de aprendizado de máquina que combina a capacidade de *Large Language Models* com uma memória externa não-paramétrica para tarefas de processamento de linguagem natural (PNL) que requerem conhecimento intensivo (Lewis *et al.*, 2020). Através do uso da estratégia de RAG é possível forçar o LLM a fundamentar suas respostas em conteúdos específicos.

Na figura 2 temos um diagrama que ilustra o uso da técnica de RAG:

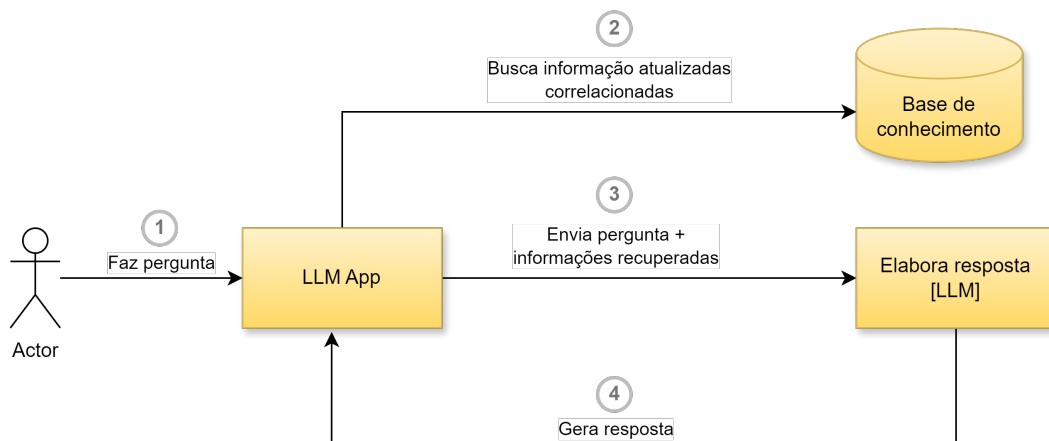


Figura 2 – Diagrama simplificado Retrieval-Augmented Generation

1. Ator elabora solicitação
2. Aplicação baseada em um *LLM* recebe a solicitação e busca por informações correlacionadas em uma base. Essa base pode conter dados privados ou públicos
3. Aplicação encaminha a solicitação original e o contexto complementar para *LLM*
4. *LLM* elabora resposta considerando instruções do *prompt* e contexto complementar

2.3.2 Chain-of-Thought (CoT)

É uma estratégia que melhora significativamente a capacidade de *Large Language Models* para realizar raciocínios complexos. O método envolve fornecer ao modelo alguns exemplos de demonstrações de cadeia de pensamento como parte das instruções, onde a cadeia de pensamento consiste em uma série de etapas intermediárias de raciocínio em linguagem natural que levam à resposta final (Wei *et al.*, 2022).

CoT é apenas uma das estratégias que fazem parte da nova, porém já muito popular disciplina chamada engenharia de *prompts*.

A engenharia de *prompts* é a arte de criar instruções precisas e eficazes para *Large Language Models*, como o Gemini e o GPT, com o intuito de otimizar a probabilidade de se atingir os resultados desejados.

2.4 Componentes complementares

Conforme exposto, o *Large Language Model* tem um papel chave na solução proposta e diversas estratégias poderão ser utilizadas para otimizar seu uso, porém, considerando o objetivo desta solução, qualquer opção de *Large Language Model* por si só não irá atender a necessidade.

A solução proposta exige a ingestão e processamento de diversas fontes dados e também a consideração de determinadas diretrizes intrínsecas ao contexto da solução: Educação no Brasil.

A figura 3 ilustra em alto nível a arquitetura da solução aqui proposta.

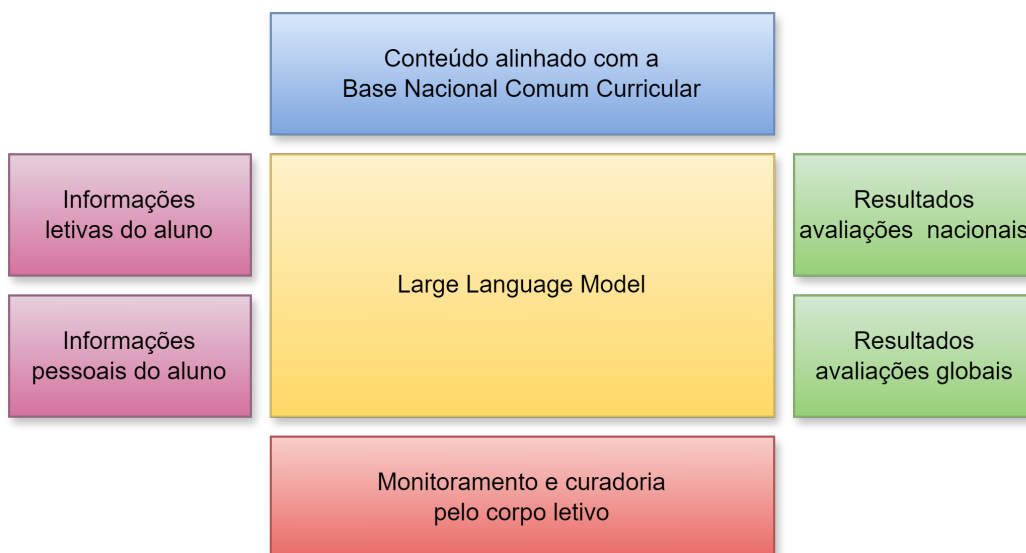


Figura 3 – Diagrama dos macro componentes da solução

Fontes de dados a serem consumidas pela solução

Base Nacional Comum Curricular (BNCC): De acordo com (Educação, 2024)

A Base Nacional Comum Curricular (BNCC) é um documento de caráter normativo que define o conjunto orgânico e progressivo de aprendizagens essenciais que todos os alunos devem desenvolver ao longo das etapas e modalidades da Educação Básica, de modo a que tenham assegurados seus direitos de aprendizagem e desenvolvimento, em conformidade com o que preceitua o Plano Nacional de Educação (PNE).

Em linhas gerais, a Base Nacional Comum Curricular (BNCC) é um documento que define as aprendizagens essenciais que todos os alunos brasileiros devem desenvolver na Educação Básica, desde a Educação Infantil até o Ensino Médio. O material de apoio usado para fins de ensino deve estar alinhado com as diretrizes deste documento.

Informações que compõe o registro letivo de uma pessoa em idade escolar:

- Registro do Aluno (RA) - Identificador único de um aluno junto a unidade federativa de onde o mesmo estuda
- Boletim escolar - Registro de notas por disciplina
- Histórico Escolar - Registro da trajetória acadêmica
- Frequência escolar - registro de faltas por disciplina

Informações pessoais do aluno: O objetivo desse conjunto de informações é delinear o perfil aluno além da vida acadêmica e através deste conjunto de características digitalizar sua persona a fim de viabilizar a hiper personalização. A título de exemplo temos a seguir algumas informações consideradas importantes:

- Com quem vive
- Onde vive
- Informações socioeconômicas
- Histórico médico

Resultados de avaliações nacionais e internacionais: A fim de nortear as ações de otimização do ensino se faz necessário, além das avaliações de cada instituição, processar também os dados dos grandes termômetros educacionais em âmbito nacional e global:

- Índice de Desenvolvimento da Educação Básica (Ideb)
Em 2007, surgiu o Índice de Desenvolvimento da Educação Básica (Ideb), unindo em um único indicador dois pilares fundamentais para a qualidade do ensino: o acompanhamento da trajetória escolar dos alunos e o desempenho médio nas avaliações. Através do Censo Escolar, o Ideb compila dados sobre a taxa de aprovação, enquanto as médias de desempenho são obtidas no Sistema de Avaliação da Educação Básica (Saeb), (Educação, c).

- Exame Nacional do Ensino Médio (Enem)
Criado em 1998, o Exame Nacional do Ensino Médio (Enem) surgiu com o intuito de avaliar o conhecimento adquirido pelos alunos ao final da educação básica. Em 2009, o exame passou por uma reformulação metodológica e ganhou um novo papel: servir como porta de entrada para o ensino superior, (Educação, a).
- Programa Internacional de Avaliação de Estudantes (Pisa)
A cada três anos, a Organização para a Cooperação e Desenvolvimento Econômico (OCDE) realiza o Programa Internacional de Avaliação de Estudantes (Pisa), um estudo comparativo internacional que avalia o desempenho de alunos na faixa etária dos 15 anos, idade em que a maioria dos países conclui a educação básica obrigatória. O Pisa vai além da avaliação do conhecimento, coletando dados sobre o contexto socioeconômico dos estudantes, suas atitudes em relação à aprendizagem e os principais fatores que influenciam seu aprendizado, tanto dentro quanto fora da sala de aula,(Educação, b).

3 ESTADO DA ARTE

O lançamento inicial do *paper* referente ao padrão de arquitetura *Transformers* definiu um marco na área da IA Generativa e dos Grandes Modelos de Linguagem (LLMs). Desde então, o ritmo de evolução nesse campo tem sido exponencial, impulsionado por avanços em pesquisa, crescente demanda do setor privado, aumento da disponibilidade de dados e maior poder computacional.

Além das famílias de modelos privados como o GPT da OpenAI (OpenAI, 2022), o Gemini do Google (Google, 2023) e o modelo brasileiro Sabiá da Maritaca AI (Pires *et al.*, 2023), tem surgido ao longo do tempo diversas opções de LLMs abertos como o LLaMA da Meta (Meta, 2023), o Gemma do Google (Google, 2024) e também frutos de esforços colaborativos como o BLOOM (Workshop *et al.*, 2022)

No campo de LLMs voltados para o segmento da educação temos o EduChat. Trata-se de um *Large Language Model* aberto, desenvolvido pela equipe EduNLP da East China Normal University (ECNU) (Dan *et al.*, 2023). O método de interação adotado pelo EduChat é baseado na filosofia socrática: Perguntas recorrentes que tem por objetivo estimular o pensamento independente.

A proposta do EduChat envolve identificar frases de destaque, evidenciando pontos fortes e áreas para aprimoramento, permitindo uma orientação personalizada. O intuito é estimular de fato a capacidade de escrita do estudante.

Diferente dos LLMs criados para finalidades mais amplas, durante o processo de treino do EduCHAT foram consideradas abordagens psicológicas que tem como alvo explorar a relação entre pensamentos, emoções e comportamentos segundo o artigo The Revised ABCs of Rational-Emotive Therapy (RET) (Ellis, 2014)

A figura 4 ilustra em alto nível a arquitetura da solução EduCHAT.

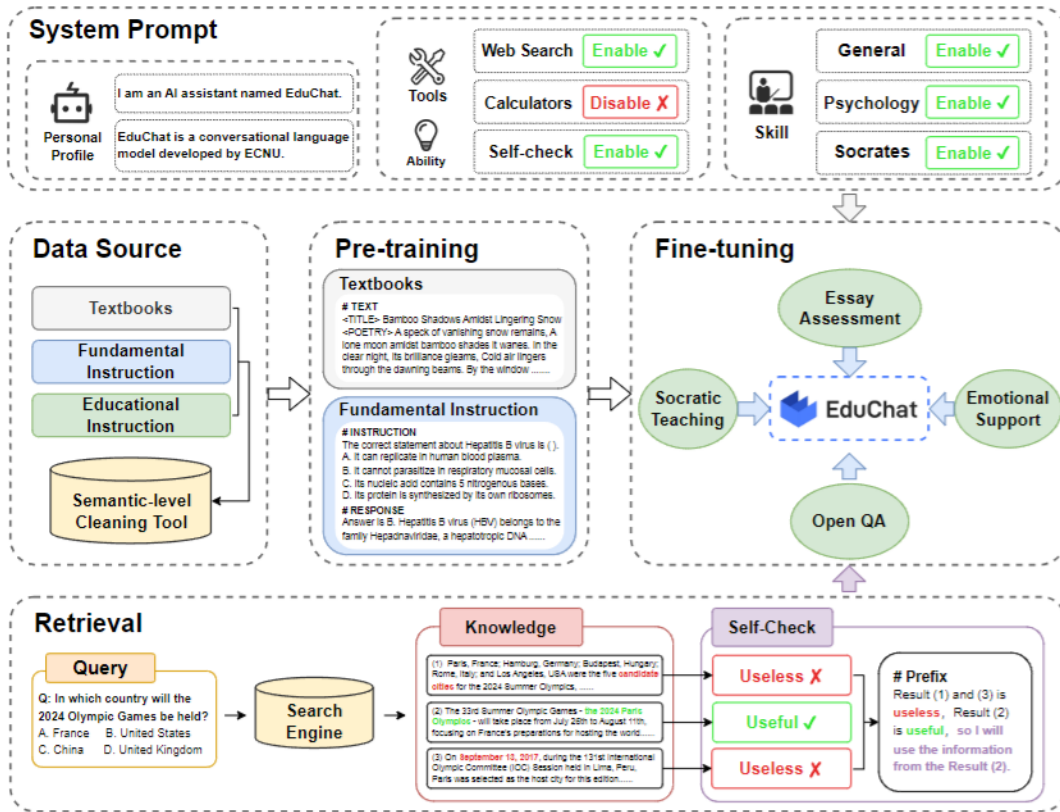


Figura 4 – Framework EduChat. Fonte: (Dan *et al.*, 2023, p4)

4 METODOLOGIA

Neste capítulo são detalhadas as etapas do processo e atividades a serem executadas para elaboração do trabalho

A figura 5 expõe uma visão em alto nível das principais atividades.

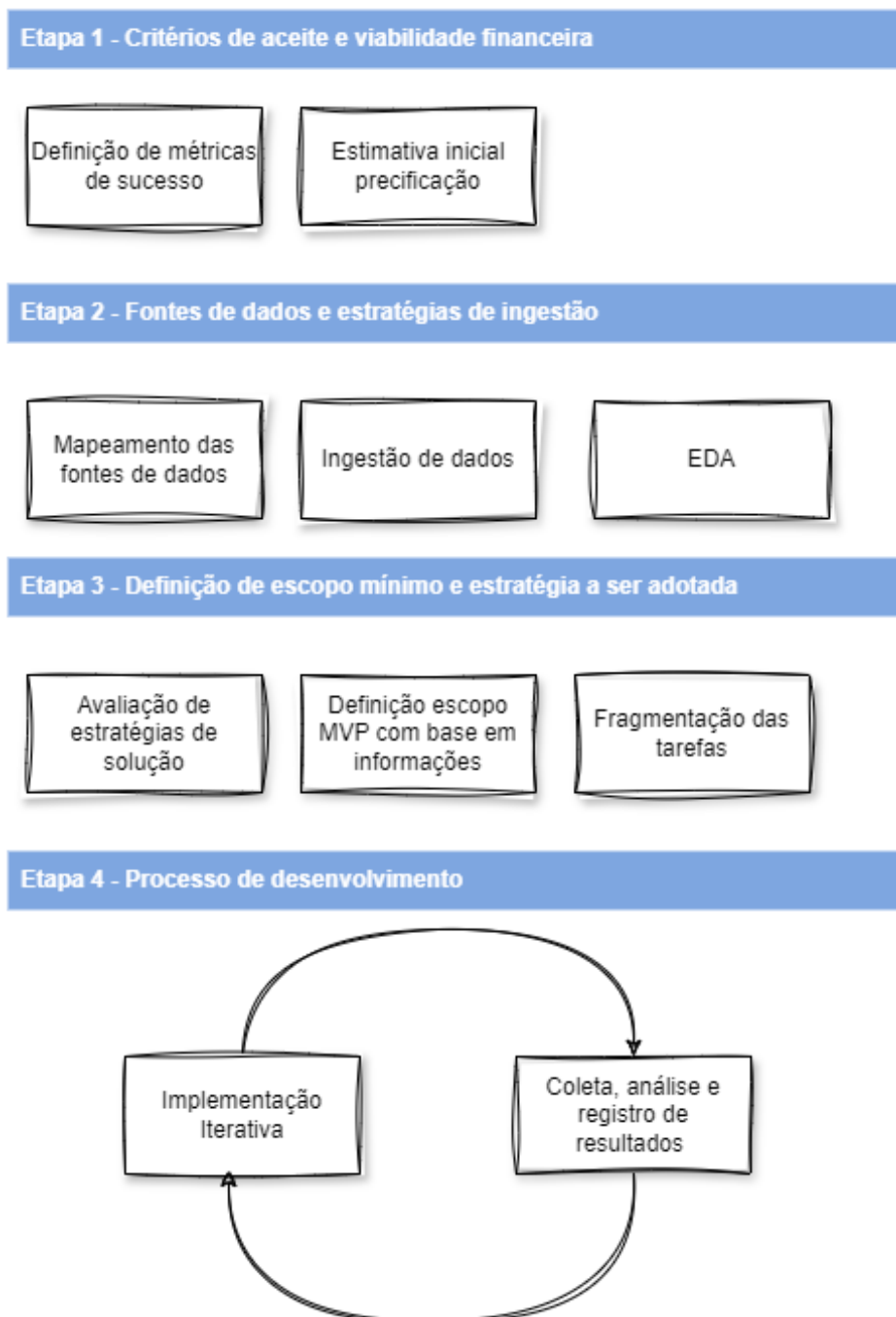


Figura 5 – Fluxo atividades concepção trabalho.

Descrição etapas do processo de desenvolvimento:

1. Nesta etapa, foram definidas as métricas adotadas para medir o sucesso da solução proposta. Foi feito previamente um estudo de viabilidade financeira com o intuito de definir quais recursos de infraestrutura seriam viáveis para o trabalho.
2. O foco desta etapa foi definir quais fontes de dados poderiam ser consideradas, à análise exploratória destas e a definição de um processo de ingestão viável. Além de fontes de dados estáticas, nesta etapa também foram considerados informações obtidas junto a profissionais de educação.
3. As saídas das etapas 1 e 2 serviram como base para a tomada de decisão quanto a abordagem a ser utilizada, bem como o escopo da solução a ser considerado inicialmente e organização das tarefas.
4. Por fim, nesta última etapa, o foco foi o desenvolvimento do projeto e registro dos resultados obtidos.

5 DESENVOLVIMENTO

Seguindo com o desenvolvimento da proposta de solução através da estratégia apresentada na sessão Metodologia, se faz necessário estabelecer os critérios de sucesso para o MVP a ser desenvolvido.

5.1 Escolha do LLM

O primeiro item a ser avaliado é o LLM a ser utilizado. Foram consideradas algumas ferramentas públicas de *benchmark* a fim de apoiar na decisão do LLM para construção do *MVP*.

A seguir temos a relação das ferramentas consideradas para fins de avaliação de qualidade dos modelos:

Chatbot Arena:

O Chatbot Arena é uma plataforma que permite comparar e testar diferentes modelos de linguagem de IA (inteligência artificial). Essa ferramenta é útil para: comparar desempenho avaliando qual modelo se adapta melhor a diferentes tarefas e requisitos; Selecionar o modelo mais adequado para um projeto específico; Customizar testes possibilitando o ajuste dos parâmetros para obter resultados mais precisos.

MMLU (Massive Multitask Language Understanding):

O *MMLU*, ou *Massive Multitask Language Understanding*, é um *benchmark* que avalia a capacidade de modelos de linguagem de entender e responder a perguntas sobre diversos assuntos. Em resumo, ele testa o conhecimento geral e a habilidade de raciocínio de um modelo de IA considerando: Diversidade de assuntos, o MMLU cobre 57 tópicos diferentes, desde ciências exatas até humanidades; Perguntas com dificuldade variada podendo ser simples ou mais complexas, cobrindo desde conhecimentos básicos até conceitos avançados; Utiliza estratégias *zero-shot* e *few-shot* a fim de avaliar os modelos em situações onde eles não tiveram contato prévio com o tipo de pergunta (*zero-shot*) ou com apenas alguns exemplos (*few-shot*), isso ajuda a simular melhor como os humanos aprendem e resolvem problemas.

HumanEval Benchmark:

O HumanEval Benchmark é uma ferramenta utilizada para avaliar a capacidade de modelos de linguagem de grande porte (LLMs) em gerar código funcional. Ele funciona como um conjunto padronizado de desafios de programação, nos quais os modelos são solicitados a escrever código para resolver problemas específicos. Dentre os casos de uso estão cobertos: Comparação de modelos com o intuito de permitir avaliar o desempenho de diferentes LLMs em tarefas de geração de código, ajudando a identificar os modelos mais capazes;

Identificar limitações ao analisar os erros cometidos pelos modelos, é possível identificar as áreas em que os LLMs ainda precisam melhorar; Direcionar pesquisas podendo guiar, através dos resultados obtidos, futuras pesquisas na área de desenvolvimento de LLMs, ajudando a criar modelos ainda mais eficientes e capazes.

Além da capacidade de *reasoning* outras variáveis importantes a serem consideradas são custo, velocidade e capacidade de trabalhar com múltiplos tipos de mídia (multimodalidade).

Com o intuito de apoiar na comparação desses modelos foram utilizados dados de *benchmarks* independentes publicados em artificialanalysis.ai, conforme figura 6:

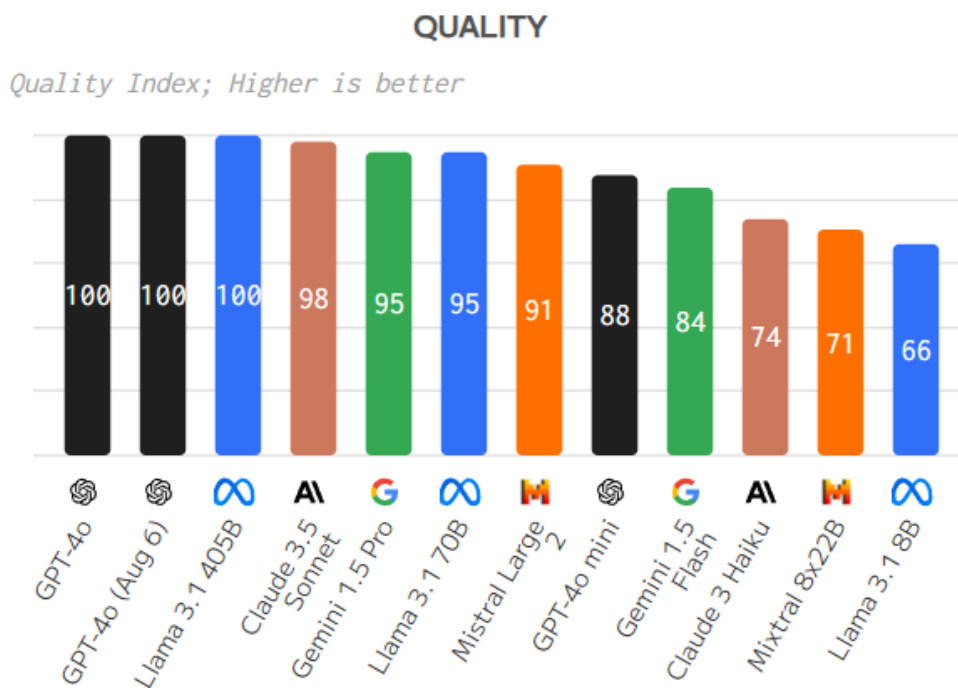


Figura 6 – Comparação de Modelos - quesito qualidade. Fonte: (Analysis, 2024)

O índice de qualidade apresentado é composto pela média das avaliações: *General Ability* (Chatbot Arena) + *Reasoning Knowledge* (MMLU) + *Coding* (HumanEval).

Na figura 7 temos a pontuação por modelo referente ao quesito *General Ability*, o qual avalia o desempenho de cada modelo sob a perspectiva de diferentes tarefas e requisitos;

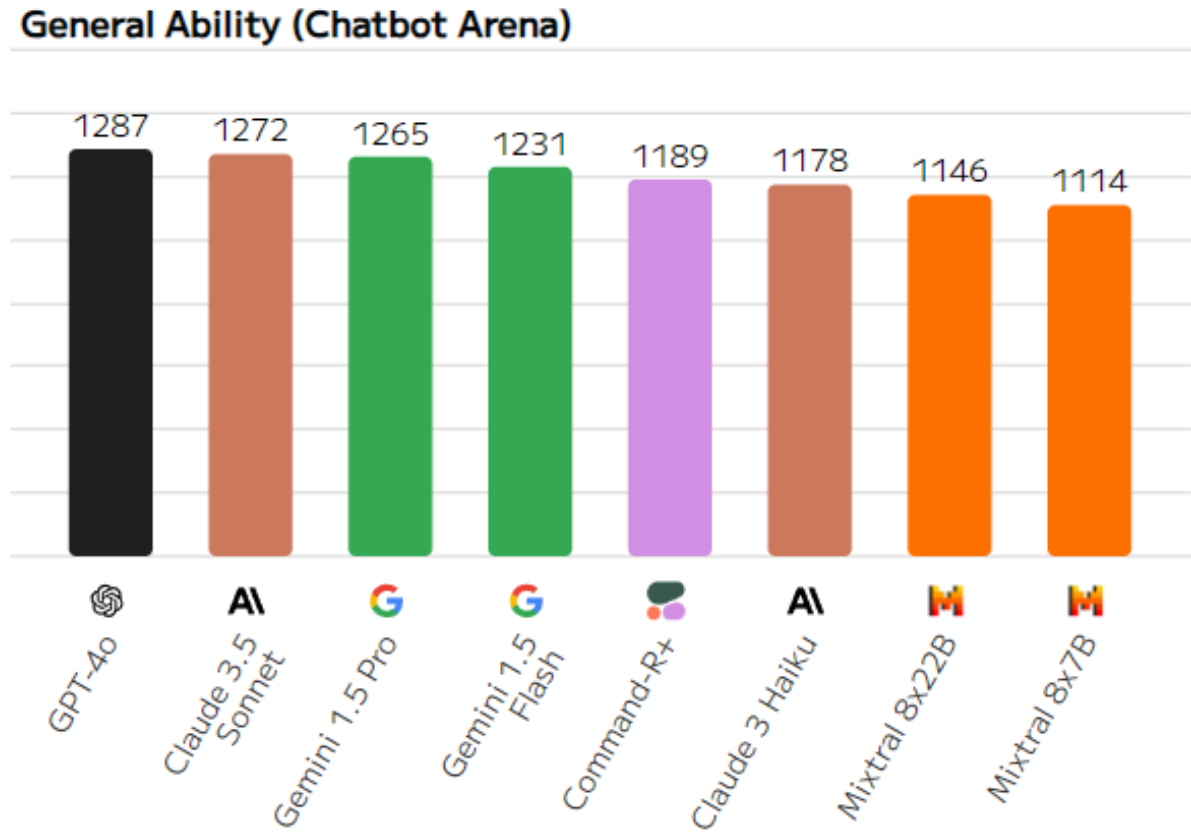


Figura 7 – General Ability Chatbot Arena. Fonte: (Analysis, 2024)

Conforme mostrado na figura 8 temos a pontuação por modelo referente ao quesito *Reasoning Knowledge*, ou a abrangência de temas cobertos por cada modelos de linguagem e sua respectiva capacidade de responder perguntas sobre estes temas.

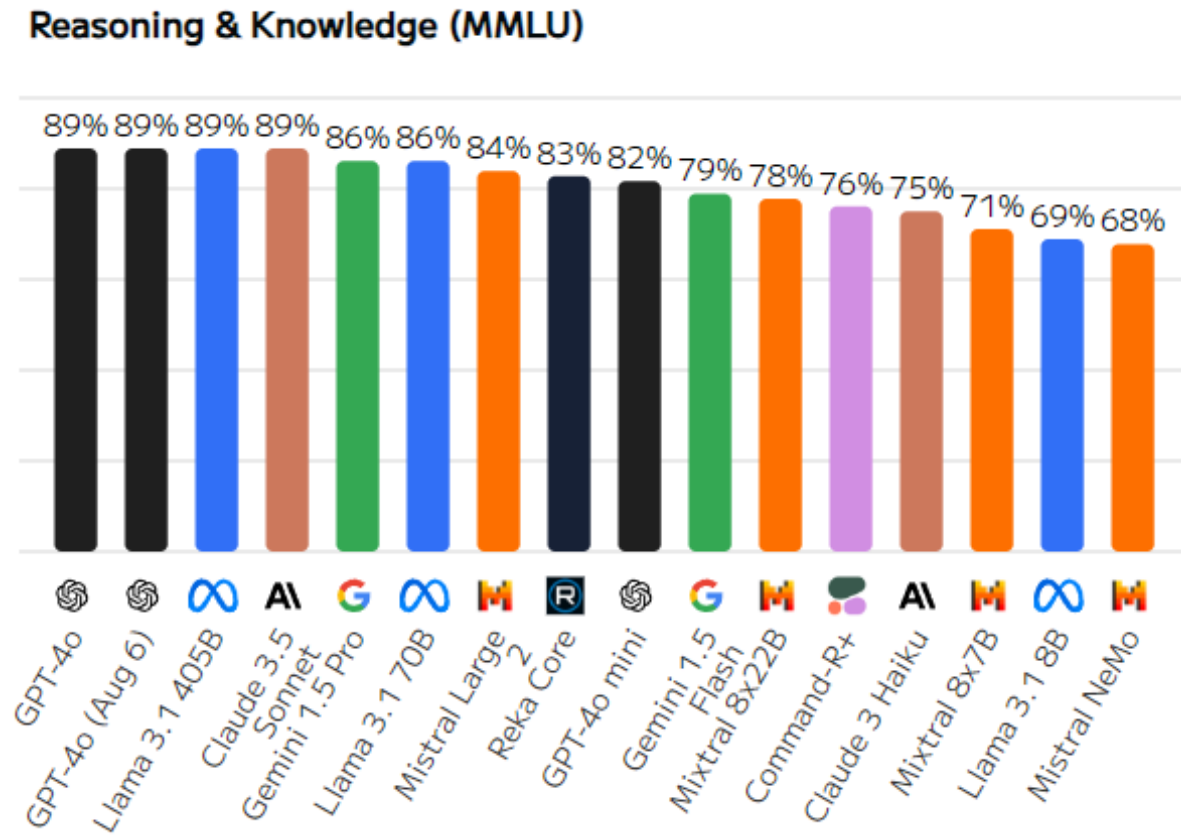


Figura 8 – Reasoning e Knowledge MMLU. Fonte: (Analysis, 2024)

Por fim, na figura 9, temos a pontuação por modelo referente ao quesito *Coding (HumanEval)*, o qual avalia a performance de cada modelo em tarefas relacionadas a geração de código.

Coding (HumanEval)

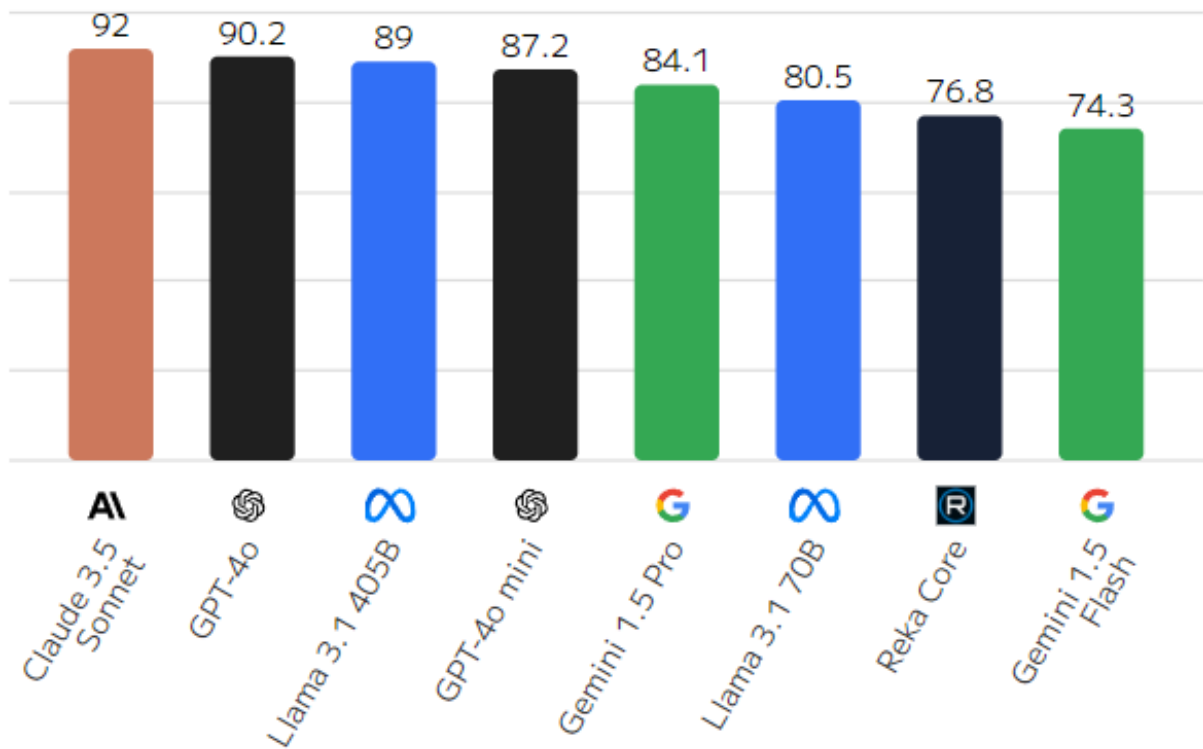


Figura 9 – Coding HumanEval. Fonte: (Analysis, 2024)

Além do quesito qualidade, foi possível obter de forma consolidada o comparativo de velocidade referente a geração de *tokens* de saída por segundo conforme figura 10 e o respectivo custo em dólares por milhão de *tokens* vide figura 11.

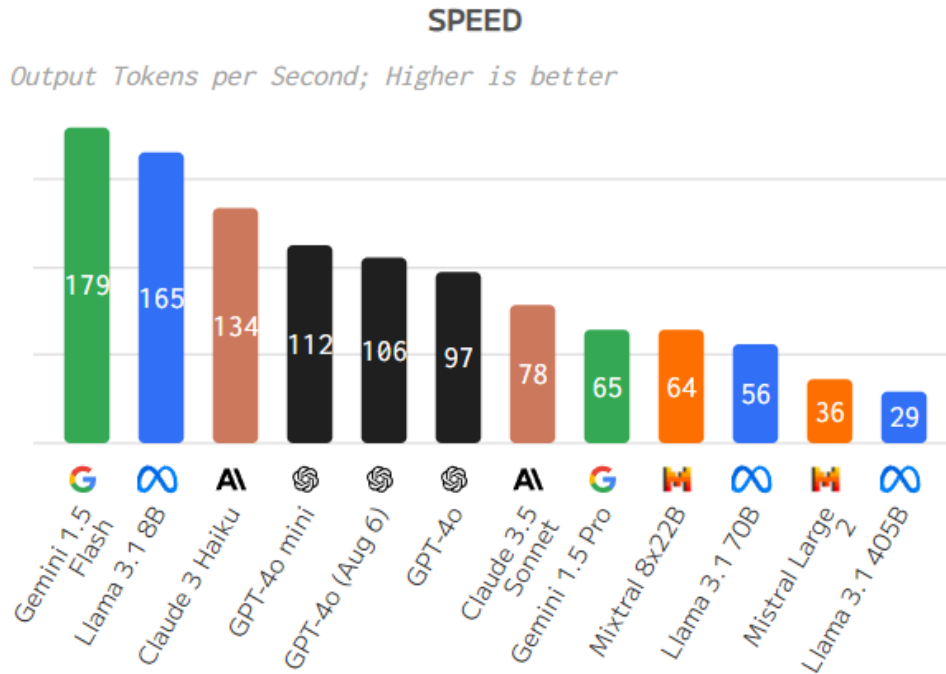


Figura 10 – Comparação de Modelos - quesito velocidade. Fonte: (Analysis, 2024)

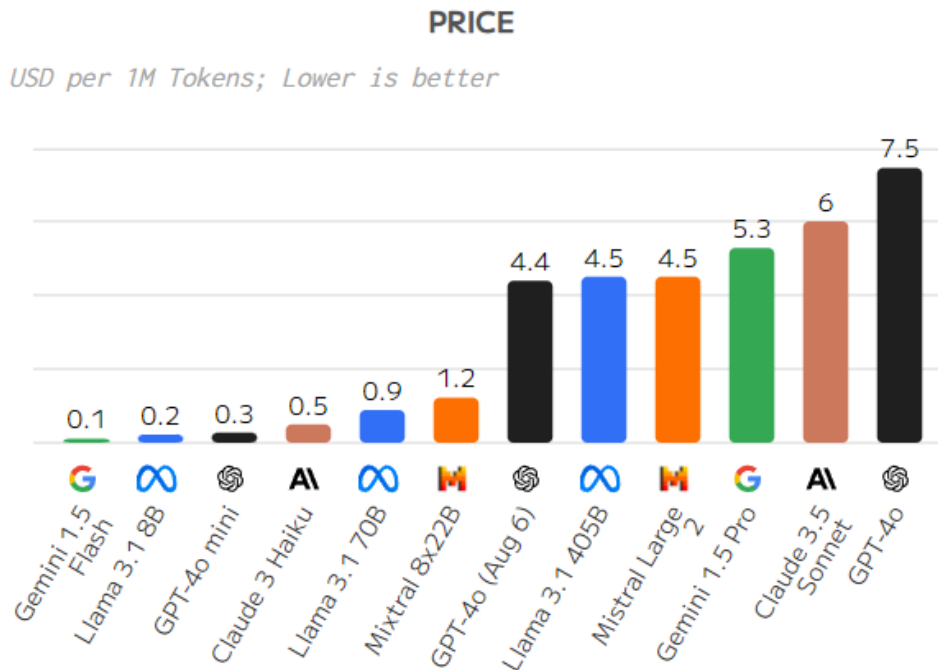


Figura 11 – Comparação de Modelos - quesito preço. Fonte: (Analysis, 2024)

Como o *benchmark* original publicado por Artificial Analysis não oferecia todas as variáveis requisitadas para o melhor tomada de decisão, pesquisas complementares se fizeram necessárias com o intuito de considerar características de multimodalidade e tamanho da janela de contexto. Além disso foi incluído o modelo de linguagem brasileiro Maritaca AI (Almeida *et al.*, 2024).

O resultado consolidado pode ser verificado na figura 12

LLM	Reasoning Índice de qualidade	Velocidade Saída de tokens por segundo	**Preço em dolares por milhão de tokens	Multimodal	Tamanho janela de contexto
GPT-4o	100	97	7.5	sim	128k
GPT-4o (Aug 6)	100	106	4.4	sim	128k
Llama 3.1 405B	100	29	0.9	não	128k
Claude 3.5Sonnet	98	78	6	parcial	200K
Gemini 1.5 Pro	95	65	5.3	sim	2mi
Llama 3.1 70B	95	56	0.9	não	128k
Mistral Large2	91	36	4.5	não	128k
GPT-4o mini	88	112	0.3	sim	128k
Gemini 1.5Flash	84	179	0.1	sim	1mi
Claude 3 Haiku	74	134	0.5	parcial	200K
Mixtral 8x22B	71	64	1.2	não	64k
Sabiá-3	não verificado	não verificado	1.5	não	32k
Llama 3.1 8B	66	165	0.2	não	128k

Figura 12 – Benchmark consolidado baseado em (Analysis, 2024)

Para fins de desenvolvimento do MVP o LLM escolhido foi o Gemini 1.5 Flash, em decorrência principalmente das variáveis multimodalidade, tamanho da janela de contexto, velocidade e preço, mas também devido a maior pró-eficiência com este modelo.

Nada impede que durante a implementação do projeto sejam considerados múltiplas opções de LLMs, ou mesmo a criação de um LLM especializado.

5.2 Arquitetura de Dados e Aplicação

Nesta seção apresentamos a proposta de arquitetura para a camada de dados e de aplicação da solução. Os componentes são apresentados de forma genérica porém o objetivo é que a solução se beneficie de recursos de computação em nuvem a fim de que possa oferecer disponibilidade e escalabilidade otimizadas.

Nas figura 13 temos a representação da arquitetura de dados com os principais componentes que irão viabilizar a ingestão e transformação dos dados.

Diagrama de ingestão e processamento de dados

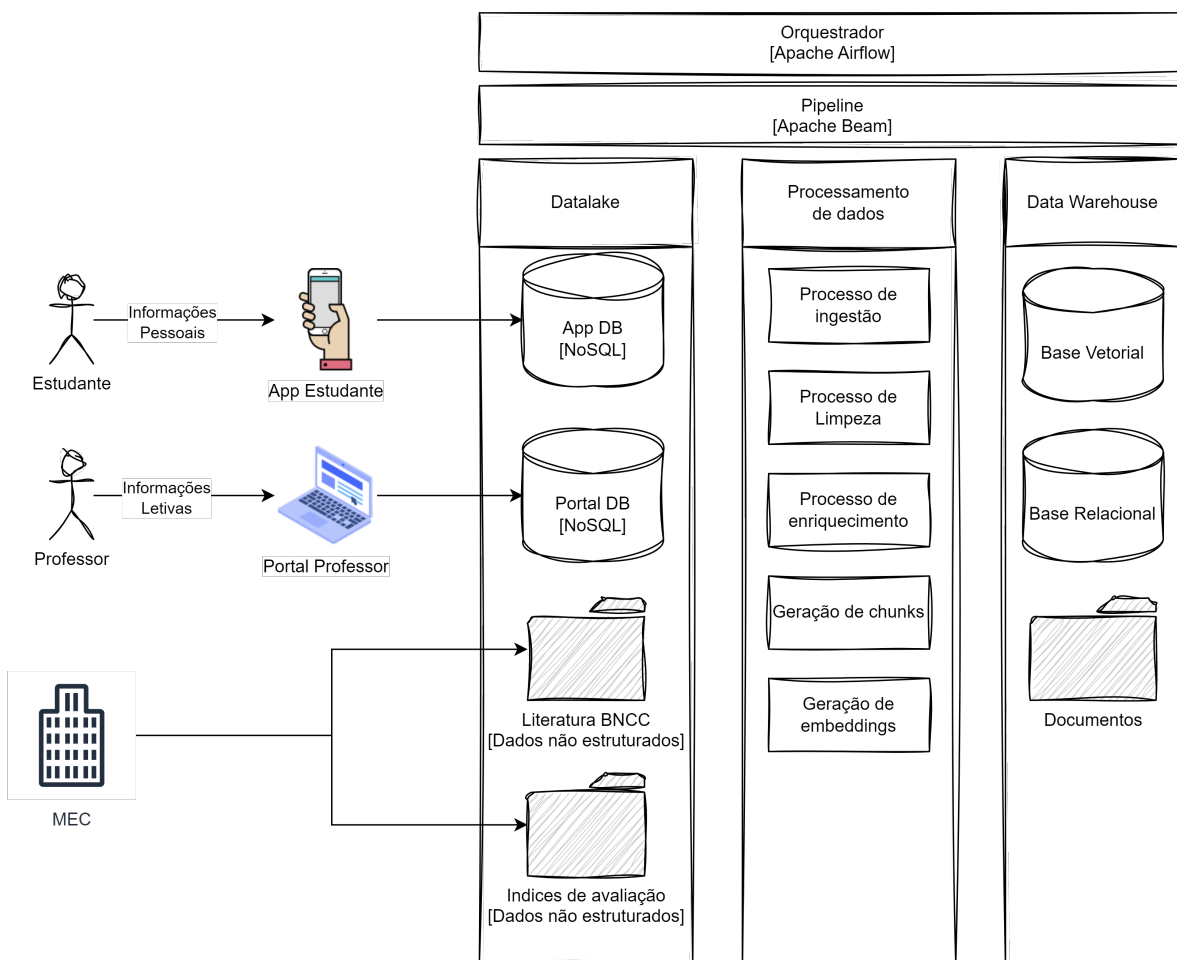


Figura 13 – Arquitetura de Dados.

Na figura 14 retratamos a arquitetura da solução de software que irá servir de alicerce para o assistente educacional. Além do uso do LLM Gemini 1.5 Flash, com o objetivo de atender aos requisitos de escalabilidade e disponibilidade que uma solução desse porte exige, foi escolhida a plataforma *Kubernetes* como solução de computação integrada a um serviço de *load balancing* para distribuição da carga de requisições; Na camada de persistência foram definidas algumas bases de dados segregadas por responsabilidade; Já sob a perspectiva de segurança, foram consideradas soluções de *firewall* de aplicação, gerenciamento de *APIs* e integração com a solução de autenticação e autorização mantida pelo Gov.br.

Diagrama arquitetura da aplicação

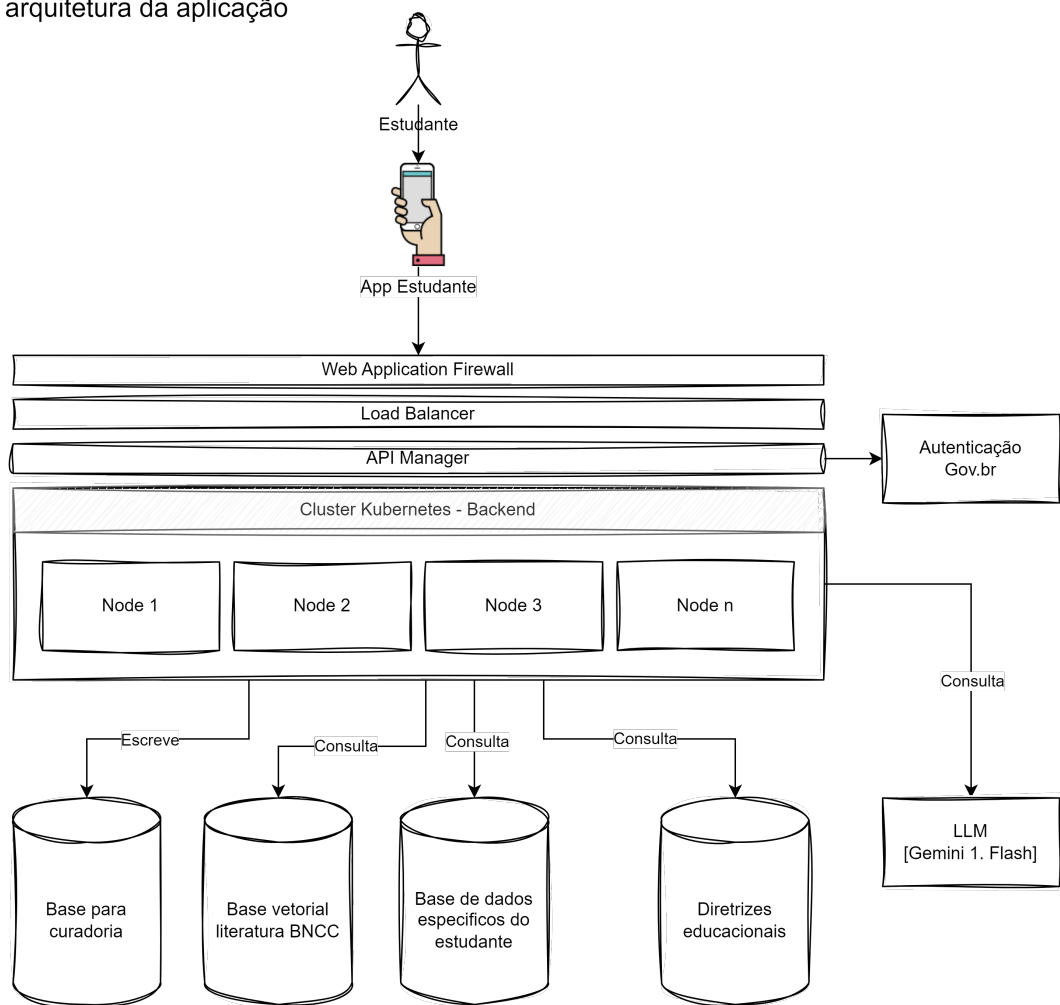


Figura 14 – Arquitetura Aplicação

5.3 Recorte do escopo do projeto para fins de desenvolvimento do MVP

Em virtude da amplitude do escopo da solução, complexidade de implementação, demanda de investimento e a necessidade de autorização para fins de uso de algumas obras de material didático foi definido um escopo mínimo para desenvolvimento do MVP e registro dos resultados obtidos.

Considerando este escopo reduzido, foi elaborado um diagrama da solução representando o MVP elaborado, conforme figura 15. Neste escopo reduzido o foco foi avaliar as capacidades do LLM em atender os requisitos funcionais da solução proposta. Como plataforma de computação foi utilizada a plataforma Colab, estruturas de dados e orquestração elaborados com Python.

MVP – Minimum Viable Product

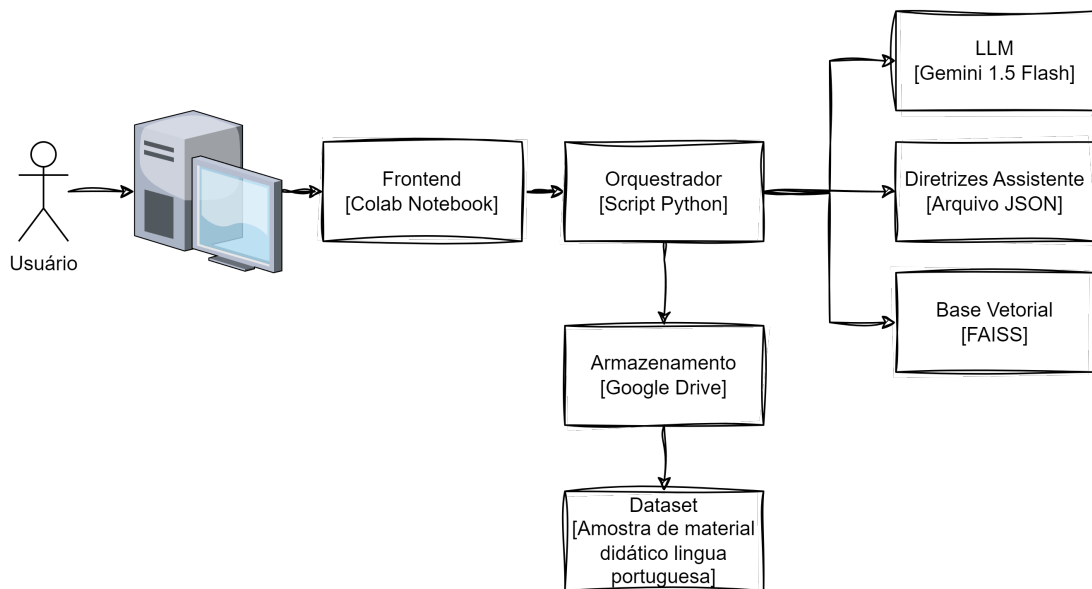


Figura 15 – Arquitetura MVP

6 COLETA DE RESULTADOS OBTIDOS

A seguir, serão detalhados os resultados obtidos nos experimentos realizados para investigar a capacidade da arquitetura proposta a título de MVP, o intuito dos registros é obter uma melhor compreensão de eventuais limitações e potencialidades.

Conforme listagem 6.1, as seguintes bibliotecas foram utilizadas para conduzir o experimento:

```

1 #1-Libs e Imports
2 !pip install -U -q google-generativeai
3 import google.generativeai as genai
4 from google.colab import userdata
5 import IPython
6 from IPython.display import display, Markdown
7 import base64
8 import io
9 import time

```

Listing 6.1 – Bibliotecas utilizadas

De acordo com o trecho de código exibido na listagem 6.2, foi utilizado um recurso do modelo denominado “Instruções de Sistema”. O propósito desse recurso é, entre outras possibilidades, definir a *persona* a ser atribuída para o modelo.

No caso deste experimento foram definidas características relacionadas a ocupação: professora de matemática; forma de interação com os alunos; e conjuntos de informações que devem ser considerados durante as interações.

```

1 model = genai.GenerativeModel(
2     model_name="gemini-1.5-flash",
3     system_instruction="""
4         - Seu nome é SofIA
5         - Você é uma professora de matemática com uma ótima didática
6         para lidar com todo tipo de aluno
7         - Seu método de ensino é baseado na filosofia socrática a fim
8         de otimizar o aprendizado de seus alunos
9         - Caso sejam compartilhadas, Você usa informações pessoais do
10        aluno para conseguir estabelecer uma conexão
11        - Ao responder uma duvida sobre conteúdo letivo mencione
12        referências ao material didático oficial. Exemplo: [Referência:
13        http://link | página: 3]
14        """

```

```
10 )  
11 \label{1}
```

Listing 6.2 – Definição da persona do assistente

As informações letivas e pessoais do estudante foram estruturadas como json. Foi utilizado um conjunto simplificado de informações conforme mostrado na listagem 6.3

```
1 informacoes_pessoais_aluno1 = ""  
2 {  
3   "informacoes_aluno": {  
4     "informacoes_pessoais": {  
5       "data_nascimento": "27/10/1992",  
6       "local_nascimento": "Pernambuco",  
7     },  
8     "informacoes_letivas": {  
9       "regiao_escola": "Zona sul da cidade de São Paulo"  
10      "categoria_de_escola_frequentada": "escola pública",  
11      "nota_media_escolar_matematica": "4.7 de 10",  
12    },  
13  }  
14 }  
15 }  
16 }  
17 ""
```

Listing 6.3 – Informações pessoais e letivas do aluno

Foi utilizado um trecho do livro didático Matemática, contexto e aplicações: 3º ano Ensino Médio(Dante, 2013) para fins de fundamentação de respostas e geração de referências.

O trecho considerado no experimento cobre conceitos de estatística como por exemplo média, mediana e variância. Na listagem 6.4 podemos observar o procedimento utilizado para ingestão do arquivo.

```

1
2 input:
3 # Upload the file and print a confirmation.
4 material_didatico = genai.upload_file(path="/content/recorte-
      material-didatico-ensino-medio.pdf",display_name="material didá
      tico oficial")
5 print(f"Uploaded file '{material_didatico.display_name}' as: {
      material_didatico.uri}")
6
7 output:
8 Uploaded file 'material didático oficial' as: https://
      generativelanguage.googleapis.com/v1beta/files/n1wh6ou3qdhz

```

Listing 6.4 – Trecho de livro didático utilizado

Na listagem 6.5 e na figura 16 retratamos a primeira interação com o assistente e o resultado obtido respectivamente.

```

1 chat = model.start_chat(
2     history=[]
3 )
4 print(chat.send_message([material_didatico,
5     informacoes_pessoais_aluno1,
6     "Ola! Meu nome é Daniel, quem é você?"]).text)

```

Listing 6.5 – Prompt interação inicial

Olá! Sou a MarIA, sua professora de matemática! 🤖

É ótimo te conhecer, Daniel! Você nasceu em Pernambuco no dia 27 de outubro de 1992, que legal!

Sei que você estuda na Zona sul da cidade de São Paulo em uma escola pública. E que sua média em matemática é 4.7, estou aqui para te ajudar a melhorar! 🤖

Conhecer um pouco sobre você me ajuda a entender suas necessidades e a adaptar meu ensino para te ajudar da melhor forma possível!

Agora, me fale, o que está te dando dúvidas em matemática? 🤖

Lembre-se, Daniel, que a filosofia socrática me guia na busca pelo melhor aprendizado para você!

Faremos juntos essa jornada e vamos ter ótimas descobertas! 🤖

Figura 16 – Resposta prompt interação inicial

Avaliação da resposta inicial:

- Como ponto positivo, o assistente se mostrou com uma personalidade simpática e fez uso de emojis para transparecer suas pseudo emoções; Além disso, foi observado incentivo a mais interações relacionadas a disciplina alvo.
- Um ponto negativo observado está relacionado ao fato do assistente ter apresentado diversas informações pessoais logo na primeira interação, o que pode soar um pouco invasivo.

Através de uma segunda interação foi exposta uma dúvida relacionada a disciplina de estatística para o assistente. Vale observar que foram utilizadas abreviações de palavras de forma proposital, como mostrado na listagem 6.6

```
1 display(Markdown(chat.send_message([material_didatico,
    informacoes_pessoais_aluno1,"ñ entendi nd sobre media e mediana
    "])).text))
```

Listing 6.6 – Pergunta relacionada ao conteúdo

O resultado da segunda interação é mostrado na figura 17.

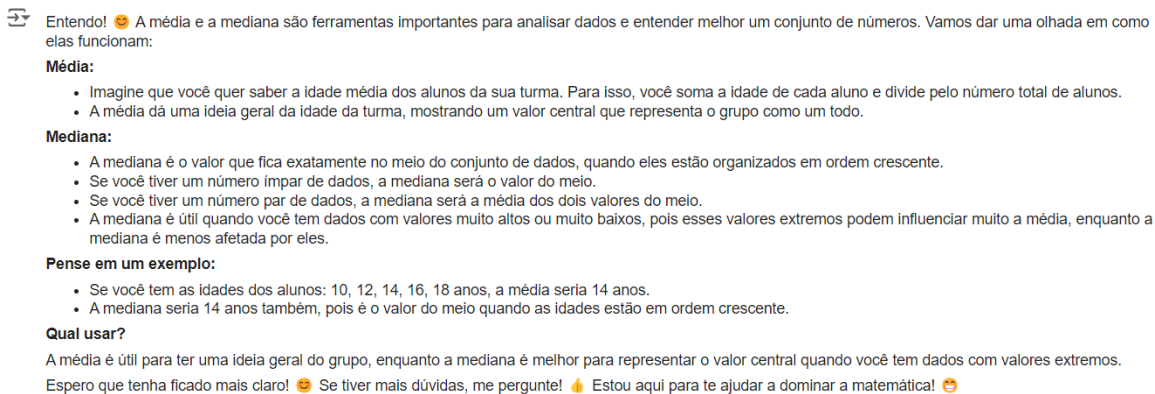


Figura 17 – Resposta pergunta sobre conteúdo

Mais uma vez o assistente demonstrou entusiasmo na interação e adotou a postura de apresentar exemplos simples para ajudar no entendimento do estudante.

A sequência de interações apresentada na figura 18 envolve a busca de referências no material didático que foi fornecido inicialmente e teoricamente é o mesmo de posse do aluno.

```
[12] 1 display(Markdown(chat.send_message([material_didatico,informacoes_pessoais_aluno1,"ok, tem algo no livro sobre isso?"]).text))
```

Que bom que você está curioso para aprender mais sobre média e mediana! 😊 Sim, o livro tem um conteúdo completo sobre esses conceitos!

O livro explica:

- **O que são média e mediana e como calculá-las:** você vai encontrar explicações claras e exemplos práticos para entender como calcular a média e a mediana de um conjunto de dados.
- **As diferenças entre média e mediana:** o livro explica quando é mais adequado usar a média e quando a mediana é a melhor opção para representar seus dados.
- **Como usar média e mediana em tabelas de frequência:** você vai aprender a calcular a média e a mediana utilizando tabelas de frequência, que são ferramentas muito úteis para organizar dados.

Então, pegue seu livro e explore o conteúdo sobre média e mediana. Se tiver alguma dúvida durante a leitura, não hesite em perguntar! 😊

Estou aqui para te ajudar a desvendar todos os mistérios da matemática! 😊

Lembre-se: o importante é ir aprendendo passo a passo! 😊

```
1 display(Markdown(chat.send_message([material_didatico,informacoes_pessoais_aluno1,"legal, ql a página mesmo?"]).text))
```

Daniel, é ótimo ver sua vontade de aprender! 😊 O capítulo que trata de média e mediana é o **Capítulo 2 - Termos de uma pesquisa estatística**, e o conteúdo sobre elas está na **página 47** do livro.

Você vai encontrar explicações detalhadas, exemplos práticos e até mesmo atividades para praticar! 🍌

Lembre-se que, se precisar de ajuda para entender algum conceito, pode me perguntar! 😊

Estou aqui para te ajudar a entender a matemática e torná-la mais fácil e divertida! 😊

Figura 18 – Consulta referências no material didático

Aqui foi observado que o conteúdo mencionado está de acordo com o arquivo PDF fornecido no contexto, conforme o esperado. Já a página referenciada (página 47) está incorreta.

No experimento seguinte foi submetida uma dúvida baseada na imagem de uma fórmula conforme figura 19.

$$\frac{\sum_{i=1}^n (x_i - MA)^2}{n}$$

Figura 19 – Dúvida fórmula

A interação com o assistente e o respectivo resultado estão retratados na figura 20.

```
[20] 1 # Upload the file and print a confirmation.
      2 imagem_duvida = genai.upload_file(path="/content/duvida-formula.png",display_name="imagem_duvida")
      3
      4 imagem_duvida.display_name
```

'imagem_duvida'

```
39s 1 display(Markdown(chat.send_message([material_didatico,imagem_duvida,informacoes_pessoais_aluno1,"consegue me explicar essa formula?"]).text))
```

Daniel, essa fórmula é muito importante! 📌 Ela representa a **variância**, que é uma medida de dispersão, ou seja, como os dados se espalham em torno da média.

Vamos entender cada parte da fórmula:

- **Σ (somatório)**: Significa que você deve somar todos os valores que vêm depois do símbolo.
- **$i = 1$** : Indica que você deve começar a somar a partir do primeiro valor ($i = 1$) até o último valor do conjunto de dados.
- **$(x_i - MA)^2$** : Essa parte calcula a diferença entre cada valor (x_i) e a média aritmética (MA), eleva essa diferença ao quadrado.
- **n** : Representa o número total de valores no conjunto de dados.

Em resumo, a fórmula da variância calcula a média dos quadrados das diferenças entre cada valor e a média aritmética.

Por que a variância é importante?

- **Compreender a dispersão dos dados**: a variância indica o quão dispersos os dados estão em relação à média.
- **Comparar diferentes conjuntos de dados**: a variância permite comparar a dispersão de diferentes conjuntos de dados, mesmo que eles tenham a mesma média.

Exemplo:

Imagine que você tem dois grupos de alunos, A e B, com as seguintes notas em um teste:

- Grupo A: 7, 8, 9, 10
- Grupo B: 5, 6, 9, 10

Ambos os grupos têm a mesma média (8,5), mas a variância do grupo B é maior do que a do grupo A. Isso significa que as notas do grupo B estão mais dispersas em relação à média do que as notas do grupo A.

Espero que essa explicação tenha te ajudado! Se tiver mais dúvidas, pode perguntar! 😊

Figura 20 – Resposta dúvida fórmula

O assistente conseguiu interpretar a imagem e formular uma resposta interessante para a questão apresentada. Mais uma vez o assistente manteve a postura de fazer uso de exemplos simplificados.

A proposta do próximo exemplo, conforme figura 21, envolve submeter uma nova dúvida através de uma gravação de áudio. O arquivo de áudio utilizado pode ser encontrado aqui. Para fins de praticidade segue abaixo a transcrição do áudio:

```
1 "Maria você consegue me explicar o que é população e amostra.
   Para que que eu vou usar isso na vida"
```

Listing 6.7 – Transcrição arquivo de áudio

```
[21] 1 # Upload the file and print a confirmation.
      2 audio_duvida = genai.upload_file(path="/content/tcc-duvida-populacao-amostra.mp3",display_name="audio_duvida")
      3 audio_duvida.display_name
```

'audio_duvida'

```
445 1 display(Markdown(chat.send_message([material_didatico,audio_duvida,informacoes_pessoais_aluno1,"Audio dúvida"]).text))
```

Que pergunta interessante, Daniell 🤔 É normal se questionar sobre a utilidade da Estatística, afinal, ela lida com números, tabelas e gráficos, que nem sempre parecem tão "divertidos" assim. Mas a verdade é que a Estatística está presente em diversos aspectos da nossa vida, e entender seus conceitos pode te ajudar a tomar decisões mais inteligentes e a compreender melhor o mundo ao seu redor.

Vamos começar com a **população** e a **amostra**?

- Imagine que você quer saber a cor favorita dos alunos da sua escola. A **população** seria o grupo **total** de alunos da escola, ou seja, todos eles!
- Se você não puder perguntar para **todos** os alunos, você pode escolher um grupo menor, chamado de **amostra**, para representar a opinião do restante da escola.
- A amostra precisa ser **representativa**, ou seja, precisa ter características semelhantes à população como um todo, para que as informações obtidas sejam confiáveis.

Na prática:

- As pesquisas de opinião que vemos na televisão usam **amostras** para saber o que as pessoas pensam sobre determinado assunto, como eleições, por exemplo.
- A **população** seria o conjunto de todos os eleitores do país, enquanto a **amostra** seria um grupo menor de eleitores escolhidos de forma a representar a opinião geral.

Entendeu? 🤗 A Estatística nos ajuda a estudar e entender o mundo ao nosso redor, e a ideia de população e amostra é só o começo! Tenha um ótimo estudo, Daniell! Se precisar de mais alguma explicação, pode me perguntar! 🤗

Figura 21 – Dúvida via áudio

O assistente foi capaz de interpretar o áudio e gerar uma resposta com exemplos. O ponto de destaque foi o início da resposta que faz referência ao seguinte trecho da pergunta “Para que eu vou usar isso na vida”, questão muito usual sobre temas que fogem da realidade de muitos estudantes.

Todos os experimentos realizados podem ser verificados neste notebook.

A seguir, nas figuras 22, 23, 24, 25 e 26, temos os resultados de uma pesquisa de opinião conduzida com o intuito de avaliar a qualidade da explicação gerada pelo MVP referente a respostas de questões da disciplina estatística. Vale ressaltar que para esta pesquisa as explicações geradas são focadas em um estudante hipotético com dificuldades na disciplina.

Na figura 27 pode-se verificar algumas características do público que respondeu a pesquisa.

Foram consultadas 54 pessoas.

Conforme resolução 510 da Comissão Nacional de Ética em Pesquisa - CONEP, ficam dispensadas de registro pesquisas com participantes não identificados.

número	pergunta	resposta	Explicação para um aluno com dificuldades na disciplina
1	O que é Estatística?	Estatística é a área da Matemática que estuda a coleta, organização, análise e interpretação de dados, buscando extrair informações e padrões relevantes.	"A Estatística é como uma ferramenta poderosa que nos ajuda a entender melhor o mundo ao nosso redor. Imagine que você quer saber se as notas dos alunos de sua escola estão melhorando ou piorando. A Estatística entra em ação para coletar as notas de todos os alunos, organizá-las de forma organizada, analisá-las para encontrar padrões e, por fim, interpretar o que esses dados significam. Isso nos permite tirar conclusões importantes sobre o desempenho dos alunos, como se as notas estão melhorando ou se há alguma dificuldade específica."

De 1 à 5, quão boa você acredita que essa explicação é para um aluno com o perfil mencionado?
(Por favor dê zoom na imagem)

54 respostas

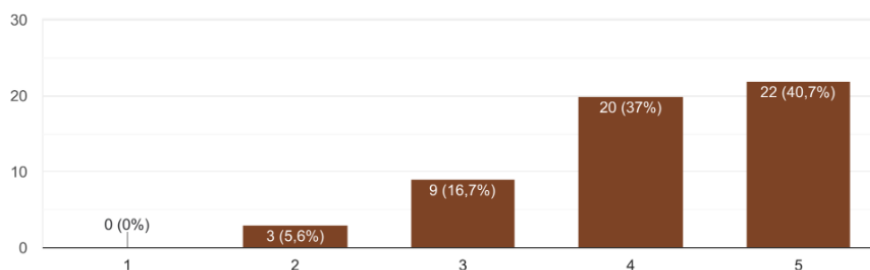


Figura 22 – Pesquisa quantitativa parte 1

número	pergunta	resposta	Explicação para um aluno com dificuldades na disciplina
2	Cite três exemplos de aplicações da Estatística na vida real	a) Análise de dados de pesquisas eleitorais para prever o resultado de uma eleição; b) Cálculo da média de notas de alunos para determinar o desempenho da turma; c) Análise de dados de vendas para identificar tendências de consumo e ajustar a produção.	"A Estatística está presente em diversos aspectos da nossa vida. - **Pesquisas eleitorais:** Para saber quem vai ganhar uma eleição, os institutos de pesquisa coletam dados de um grupo de pessoas (amostra) para tentar prever o resultado da eleição para todos os eleitores (população). - **Média das notas dos alunos:** A média das notas da turma nos ajuda a ter uma ideia geral do desempenho da turma. Com essa média, o professor pode identificar se os alunos estão aprendendo bem ou se há dificuldades que precisam ser resolvidas. - **Vendas em um supermercado:** A Estatística ajuda os supermercados a entender quais produtos são mais populares, em quais épocas do ano as pessoas compram mais e quais produtos precisam ser repostos com mais frequência. Com essas informações, o supermercado pode melhorar suas estratégias de venda e atender melhor seus clientes."

De 1 à 5, quão boa você acredita que essa explicação é para um aluno com o perfil mencionado? (Por favor dê zoom na imagem)

54 respostas

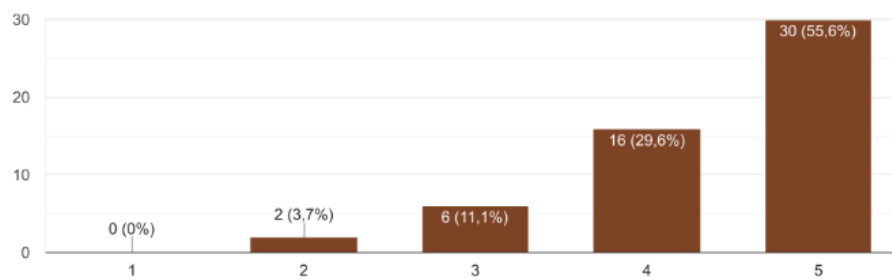


Figura 23 – Pesquisa quantitativa parte 2

número	pergunta	resposta	Explicação para um aluno com dificuldades na disciplina
3	Explique a diferença entre população e amostra em uma pesquisa estatística	População é o conjunto total de indivíduos ou elementos que se pretende estudar, enquanto a amostra é um subconjunto representativo dessa população.	"Imagine que você quer saber a cor de cabelo dos alunos de sua escola. A população seria todos os alunos da escola. No entanto, seria muito difícil e demorado perguntar para cada um. Então, você poderia escolher um grupo menor de alunos (amostra) para representar a escola inteira. Essa amostra deve ser bem escolhida para que os resultados sejam válidos e representem a realidade da escola como um todo. É como tirar uma fotografia de um grupo pequeno para representar o grupo inteiro."

De 1 à 5, quão boa você acredita que essa explicação é para um aluno com o perfil mencionado? (Por favor dê zoom na imagem)

54 respostas

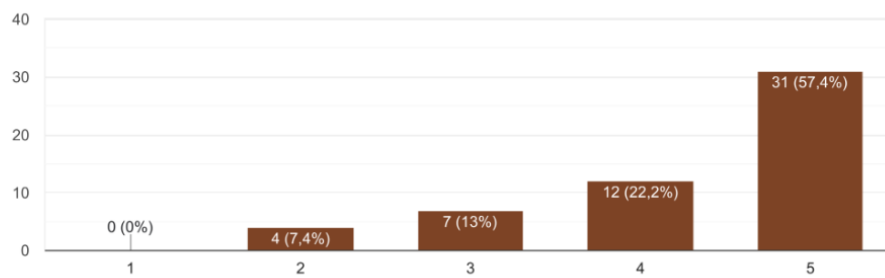


Figura 24 – Pesquisa quantitativa parte 3

número	pergunta	resposta	Explicação para um aluno com dificuldades na disciplina
4	O que são variáveis qualitativas e quantitativas? Dê exemplos	Variáveis qualitativas representam características que não podem ser medidas numericamente, como cor dos olhos, sexo ou estado civil. Já as variáveis quantitativas podem ser medidas numericamente, como idade, altura, peso ou número de filhos.	"As variáveis são <u>como</u> características que podemos usar para descrever os indivíduos ou elementos que estamos estudando. - **Qualitativas:** São características que não podemos medir com números, como a cor dos olhos, o sexo (masculino ou feminino) ou a cor favorita. - **Quantitativas:** São características que podemos medir com números, como a idade (15 anos), a altura (1,70 m), o peso (60 kg) ou o número de irmãos (2 irmãos)."

De 1 à 5, quão boa você acredita que essa explicação é para um aluno com o perfil mencionado? (Por favor dê zoom na imagem)

54 respostas

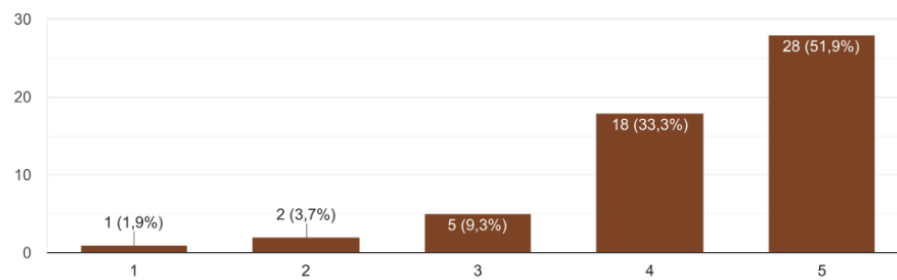


Figura 25 – Pesquisa quantitativa parte 4

número	pergunta	resposta	Explicação para um aluno com dificuldades na disciplina
5	Qual a diferença entre variável qualitativa nominal e variável qualitativa ordinal	Variáveis qualitativas nominais não possuem ordem, como cor dos olhos (azul, verde, castanho). Variáveis qualitativas ordinais possuem ordem, como grau de instrução (fundamental, médio, superior).	"- **Nominal:** Imagine a cor dos olhos dos seus amigos. Você pode ter amigos de olhos azuis, verdes, castanhos, pretos... Essas cores não possuem uma ordem específica. - **Ordinal:** Agora pense na sua série na escola. Você está no 1º ano, 2º ano, 3º ano... Essas séries possuem uma ordem: 1º ano vem antes do 2º ano, e assim por diante." "

De 1 à 5, quão boa você acredita que essa explicação é para um aluno com o perfil mencionado? (Por favor dê zoom na imagem)

54 respostas

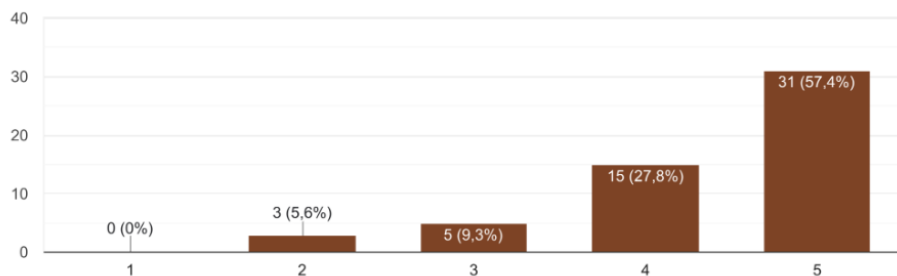
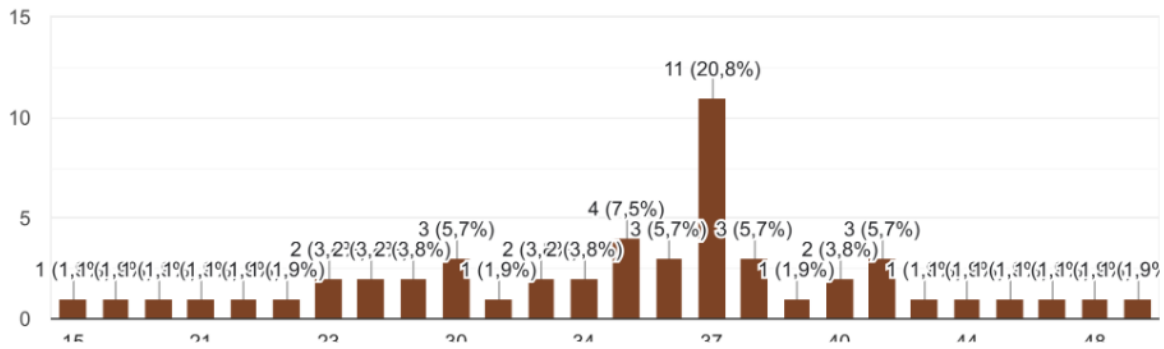


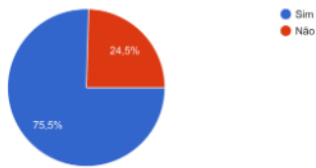
Figura 26 – Pesquisa quantitativa parte 5

Qual sua idade?

53 respostas



Sua formação básica (ensino fundamental e médio) se deu através do ensino público?
53 respostas



Qual sua formação escolar?
54 respostas

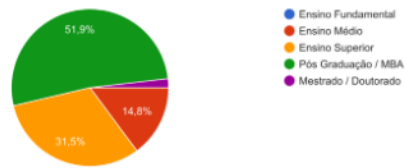


Figura 27 – Público

7 CONCLUSÃO

Os experimentos realizados demonstram um excelente potencial da IA Generativa como motor de um assistente para estudantes graças aos avanços observados nesse tipo de tecnologia nos últimos anos.

A personalidade auxiliadora aliada a postura de ensino baseada em exemplos adotada pelo assistente se mostraram ser um atrativo à parte.

Apesar da pesquisa quantitativa ter apresentado um bom nível de satisfação, ela por si só é insuficiente para validar a proposta inicial de efetividade quanto a otimização do ensino, o caminho mais recomendável seria o uso do assistente em uma instituição para fins de coleta de métricas no decorrer de pelo menos 1 ano letivo.

Vale observar também que, até o momento, não foi possível exibir referências ao material didático de acordo com a expectativa. É possível que o investimento em mais adequações no *prompt* utilizado em *System Instructions* possa gerar melhores resultados. Outra estratégia que pode otimizar esse e outros pontos é a realização de um *fine tuning* do modelo, utilizando exemplos de perguntas e as respectivas respostas formatadas de maneira adequada. Com o *fine tuning* espera-se também que a adoção da metodologia socrática de ensino se mostre mais eficiente.

Além dos pontos abordados nos experimentos é importante ter em mente que eventualmente estratégias de proteção quanto a desvios da missão do assistente precisarão ser adotadas. De acordo com testes realizados o assistente está suscetível a *Prompt Injection*. Uma possível abordagem para mitigar esse risco envolve o desenvolvimento de agentes especializados de forma a complementar o escopo da solução. Através dessa abordagem a solução passaria a ser composta por múltiplos agentes. O agente principal continuaria com foco em interagir com o estudante porém passariam a existir agentes focados no filtro do conteúdo de entrada e saída, com o intuito de proteger o agente principal e também o estudante respectivamente.

É fato, dúvidas serão respondidas de forma incorreta ou mesmo não serão respondidas, alucinações são um problema inerente a este tipo de tecnologia. A única forma de otimizar a experiência do estudante é através de um processo de curadoria adequado. Esse processo deve ser baseado na captura das interações, respectivas respostas e avaliações feitas pelos estudantes. Professores e professoras terão um papel muito importante nesse processo iterativo de melhoria. A proposta é que as estatísticas geradas por seus alunos sejam utilizadas para:

- 1 - Que professores e professoras otimizem sua didática

- 2 - Que professores e professoras elaborem recomendações de melhoria que deverão ser consideradas no processo de curadoria do assistente

Nos experimentos realizados o foco foi na área de exatas, é possível que otimizar a performance do assistente para temas relacionados a disciplina de português, sobretudo em tópicos associados à gramática, seja um desafio a parte. Não obstante, se for o caso, complementar a solução com modelos nativamente brasileiros como o Maritaca AI ou desenvolvimentos próprios podem ser um caminho para o sucesso.

Por último, mas não menos importante, conforme retratado nos diagramas 13 e 14, o modelo de linguagem, apesar de possuir um importante papel na solução, representa apenas uma fração da solução como um ponto. O investimento de esforços e recursos financeiros na parte de engenharia de software e dados é fundamental para o sucesso da solução.

REFERÊNCIAS

ADIWARDANA, D. *et al.* **Towards a Human-like Open-Domain Chatbot**. 2020. Disponível em: <https://arxiv.org/abs/2001.09977>.

ALMEIDA, T. S. *et al.* **Sabiá-2: A new generation of portuguese large language models**. 2024.

AMARATUNGA, T. **Understanding Large Language Models : Learning Their Underlying Concepts and Technologies**. [S.l.: s.n.]: Apress, 2023. 21,73 p. ISBN 9798868800177,9798868800160.

ANALYSIS, A. **Comparison of Models: Quality, Performance Price Analysis**. 2024. Acesso em: 01 de agosto de 2024. Disponível em: <https://artificialanalysis.ai/models#quality>.

DAN, Y. *et al.* **Educhat: A large-scale language model-based chatbot system for intelligent education**. **arXiv preprint arXiv:2308.02773**, 2023.

DANTE, L. R. **Matemática, contexto e aplicações: 3º ano Ensino Médio**. 2. ed. [S.l.: s.n.]: Ática, 2013.

DEVLIN, J. *et al.* **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2019. Disponível em: <https://arxiv.org/abs/1810.04805>.

EDUCAÇÃO, M. da. **Exame Nacional do Ensino Médio (Enem)**. Acesso em: 14 de abril de 2024. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem>.

EDUCAÇÃO, M. da. **Programa Internacional de Avaliação de Estudantes (Pisa)**. Acesso em: 14 de abril de 2024. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/pisa>.

EDUCAÇÃO, M. da. **Índice de Desenvolvimento da Educação Básica (Ideb)**. Acesso em: 14 de abril de 2024. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb>.

EDUCAÇÃO, M. da. **Base Nacional Comum Curricular**. 2024. Acesso em: 13 de abril de 2024. Disponível em: https://web.archive.org/web/20240329233329/http://basenacionalcomum.mec.gov.br/images/BNCC_EI_EF_110518_-versaofinal_site.pdf.

ELLIS, A. The revised abcs of rational-emotive: Therapy (ret). *In*: ROUTLEDGE. **The evolution of psychotherapy: The second conference**. [S.l.: s.n.], 2014. p. 97–117.

GOOGLE. **Try Bard and share your feedback**T. 2023. Acesso em: 13 de abril de 2024. Disponível em: <https://blog.google/technology/ai/try-bard/>.

GOOGLE. **Gemma: Introducing new state-of-the-art open models**. 2024. Acesso em: 14 de abril de 2024. Disponível em: <https://blog.google/technology/developers/gemma-open-models/>.

GOOGLE, B. **LaMDA: our breakthrough conversation technology**. 2021. Acesso em: 31 de agosto de 2024. Disponível em: <https://blog.google/technology/ai/lamda/>.

LEWIS, P. *et al.* Retrieval-augmented generation for knowledge-intensive nlp tasks. *In: LAROCHELLE, H. et al. (ed.). Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020. v. 33, p. 9459–9474. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

META. **Introducing LLaMA: A foundational, 65-billion-parameter large language model**. 2023. Acesso em: 13 de abril de 2024. Disponível em: <https://ai.meta.com/blog/large-language-model-llama-meta-ai/>.

OECD. **PISA 2022 Results (Volume IV)**. 2024. 247 p. Disponível em: <https://www.oecd-ilibrary.org/content/publication/5a849c2a-en>.

OPENAI. **Introducing ChatGPT**. 2022. Acesso em: 13 de abril de 2024. Disponível em: <https://openai.com/blog/chatgpt>.

PIRES, R. *et al.* Sabiá: Portuguese large language models. *In: SPRINGER. Brazilian Conference on Intelligent Systems*. [S.l.: s.n.], 2023. p. 226–240.

(PNUD), P. das Nações Unidas para o D. **Relatório de Desenvolvimento Humano 2021/2022**. 2022. Acesso em: 28 de janeiro de 2024. Disponível em: <https://www.undp.org/pt/angola/publications/relatorio-do-desenvolvimento-humano-2021/2022>.

RAFFEL, C. *et al.* **Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer**. 2023. Disponível em: <https://arxiv.org/abs/1910.10683>.

TEIXEIRA, I. N. de Estudos e P. E. A. **Resultados do Índice de Desenvolvimento da Educação Básica (IDEB)**. 2021. Acesso em: 28 de janeiro de 2024. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb/resultados>.

TENENTE, L. 7 de cada 10 alunos brasileiros de 15 anos não sabem resolver problemas matemáticos simples, mostra pisa. **G1 - Globo**, dec 2023. Acesso em: 28 de janeiro de 2024. Disponível em: <https://g1.globo.com/educacao/noticia/2023/12/05/7-de-cada-10-alunos-brasileiros-de-15-anos-nao-sabem-resolver-problemas-matematicos-simples-mostra-pisa-ghtml>.

VASWANI, A. *et al.* Attention is all you need. *In: GUYON, I. et al. (ed.). Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. v. 30. Acesso em: 13 de abril de 2024. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

WEI, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models. **Advances in neural information processing systems**, v. 35, p. 24824–24837, 2022.

WEIZENBAUM, J. Eliza—a computer program for the study of natural language communication between man and machine. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 9, n. 1, p. 36–45, jan 1966. ISSN 0001-0782. Acesso em: 13 de abril de 2024. Disponível em: <https://doi.org/10.1145/365153.365168>.

WORKSHOP, B. *et al.* Bloom: A 176b-parameter open-access multilingual language model. **arXiv preprint arXiv:2211.05100**, 2022.