

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

**Tutoria para o ENEM com Inteligência Artificial:
alavancando tecnologias para democratizar oportunidades**

Pedro Meyer Ditmar

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

**Tutoria para o ENEM com
Inteligência Artificial: alavancando
tecnologias para democratizar
oportunidades**

Pedro Meyer Dittmar

USP - São Carlos
2024

Pedro Meyer Dittmar

Tutoria para o ENEM com Inteligência Artificial: alavancando tecnologias para democratizar oportunidades

Trabalho de conclusão de curso apresentado 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

Orientadora: Profa. Dra. Solange O. Rezende

USP - São Carlos

2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

D617t Dittmar, Pedro Meyer
Tutoria para o ENEM com Inteligência Artificial:
alavancando tecnologias para democratizar
oportunidades / Pedro Meyer Dittmar; orientadora
Solange O. Rezende. -- São Carlos, 2024.
100 p.

Trabalho de conclusão de curso (MBA em
Inteligência Artificial e Big Data) -- Instituto de
Ciências Matemáticas e de Computação, Universidade
de São Paulo, 2024.

1. Inteligência Artificial Generativa. 2. Large
Language Models. 3. ENEM. 4. Educação. 5.
Democratização de oportunidades. I. Rezende, Solange
O. , orient. II. Título.

DEDICATÓRIA

*À inventividade do ser humano para
solucionar problemas, e ao
reconhecimento do outro como valor
para direcionar essa capacidade*

AGRADECIMENTOS

Aos professores do MBA em Inteligência Artificial e Big Data Solange O. Rezende e Ricardo Marcacini, sem o apoio de quem não teria sido possível executar este projeto.

Às pessoas que compõe a Escola Vera Cruz e a USP, e a aos tantos ótimos professores que encontrei no meu caminho, a quem devo minha formação.

Aos meus colegas de AppJusto, junto de quem eu aprendi a trabalhar com inovação e tecnologia visando o bem da sociedade.

Ao Marcelo Sodelli, que me ensina a usar a compreensão como ferramenta para superar dificuldades.

Aos meus amigos, sempre do meu lado e que pelo exemplo me fazem crer em mim mesmo.

Finalmente e principalmente, aos meus pais Maria do Carmo Vidigal e Wulf Dittmar, pelo incentivo incondicional.

EPÍGRAFE

Nenhuma tecnologia é boa ou ruim;
o que há são usos bons e ruins de tecnologias
Autor desconhecido

RESUMO

DITTMAR, P. M. **Tutoria para ENEM com Inteligência Artificial:** alavancando tecnologias para democratizar oportunidades. 2024. 100 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

Este Trabalho de Conclusão de Curso propõe o desenvolvimento de uma solução de tutoria para o Exame Nacional do Ensino Médio (ENEM), baseada em Inteligência Artificial Generativa. O objetivo é oferecer uma ferramenta acessível de preparação ao vestibular para estudantes de baixa renda, democratizando os recursos de preparação para o vestibular e ampliando oportunidades de acesso ao ensino superior. A solução combina a adaptação de Large Language Models (LLMs) do estado da arte para o contexto do ENEM com a implementação de um sistema de recomendação de questões personalizadas, fundamentado na Teoria de Resposta ao Item (TRI), visando otimizar a prática dos alunos. O trabalho discute o potencial transformacional da tecnologia para mitigar a desigualdade de oportunidades no Brasil, apresenta um modelo de negócios viável para a solução baseado no financiamento público e explora estratégias para um desenvolvimento progressivo, validado por testes com usuários reais a fim de mitigar riscos e investir recursos com eficiência.

Palavras-chave: Inteligência Artificial Generativa; Grandes Modelos de Língua; ENEM; Educação; democratização de oportunidades.

ABSTRACT

DITTMAR, P. M. **Tutoring for ENEM**: leveraging technologies to democratize opportunities. 2024. 100 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

This dissertation proposes the development of a tutoring solution for the Brazilian National High School Exam (ENEM), based on Generative Artificial Intelligence. The goal is to offer an accessible tool for low-income students to prepare for the exam, democratizing preparation resources and expanding opportunities for access to higher education. The solution combines the adaptation of state-of-the-art Large Language Models (LLMs) to the ENEM context with the implementation of a personalized question recommendation system, grounded in Item Response Theory (IRT), aiming to optimize student practice. The work discusses the transformational potential of technology to mitigate inequality of opportunities in Brazil, presents a viable business model for the solution based on public funding, and explores strategies for progressive development, validated by user testing to mitigate risks and invest resources efficiently.

Keywords: Generative Artificial Intelligence; Large Language Models; ENEM; Education; democratization of opportunities.

LISTA DE ILUSTRAÇÕES

Figura 1 – Comparação de respostas geradas por GPT-4 sem e com CoT.....	28
Figura 2 – Performance do GPT-4 no ENEM 2023.....	29
Figura 3 – Hype cycle for Artificial Intelligence, 2024.....	40
Figura 4 – Investimentos globais em IA Generativa, 2019-2023.....	41
Figura 5 – Business Canvas do projeto de tutoria para ENEM com IA Generativa.....	50
Figura 6 – Criação de função para <i>Chatbot</i> a partir de API da OpenAI.....	57
Figura 7 – Testes iniciais com API da OpenAI no contexto de tutoria (...) do ENEM...	58
Figura 8 – Base de dados utilizada para validação da hipótese técnica.....	59
Figura 9 – Experiência de uso com GPT da OpenAI configurado para (...) ENEM.....	60

LISTA DE TABELAS

Tabela 1 – Resultados no ENEM 2009-2017 com <i>zero-shot</i> prompting.....	20
Tabela 2 – Resultados no ENEM 2022.....	23
Tabela 3 – Comparativo de soluções tecnológicas de preparação ao vestibular.....	46
Tabela 4 – Matriz SWOT da solução.....	60

LISTA DE ABREVIATURAS E SIGLAS

API	–	Application Programming Interface
B2B	–	Business-to-Business
B2C	–	Business-to-Consumer
B2G	–	Business-to-Government
CoT	–	Chain-of-Thought Prompting
ENEM	–	Exame Nacional do Ensino Médio
GPT	–	Generative Pre-trained Transformer
IA	–	Inteligência Artificial
ICL	–	In-Context Learning
INEP	–	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
LLM	–	Large Language Model
MVP	–	Minimum Viable Product
RAG	–	Retrieval Augmented Generation
RLHF	–	Reinforcement Learning through Human Feedback
TRI	–	Teoria de Resposta ao Item
UX	–	User-Experience

SUMÁRIO

1 INTRODUÇÃO	1
1.1 O sentido do uso de tecnologia no combate à desigualdade de oportunidades	1
1.2 A inserção do sistema educacional brasileiro nas estruturas que perpetuam desigualdades	2
1.3 Democratização no acesso ao ensino superior como alvo de políticas públicas de alto custo benefício	4
1.4 Desafios à agência dos estudantes na preparação para o vestibular e o potencial de soluções a partir de tecnologia.....	5
1.5 Proposta de uma solução de tutoria para o ENEM baseada em Large Language Models.....	7
2 REFERENCIAL TEÓRICO	11
2.1 Teorias para o uso de tecnologia computacional na educação	11
2.2 Especificando LLMs e principais técnicas para sua adaptação	14
2.2.1 Sobre a origem e principais propriedades dos LLMs	14
2.2.2 Alternativas para adaptação de LLMs a <i>use cases</i> específicos.....	16
2.2.3 Prompting: como orientar o comportamento do modelo sem treinamento	18
2.2.4 Recuperação confiável de dados de uma base através de RAG	19
2.2.5 Um ponto de atenção: as “alucinações” dos LLMs	21
2.2.6 LLMs são capazes de resolver as questões do ENEM?	22
2.3 Teoria de Resposta ao Item (TRI)	27
2.3.1 Adequação à recomendação de questões visando uma prática otimizada para estudantes individuais	27
2.3.2 Ponderações sobre a implementação de um sistema recomendador de questões do ENEM a partir da TRI	28
2.3.3 Alternativas de sistemas recomendadores.....	30
3 MODELO DE EXECUÇÃO E REFERENCIAL PRÁTICO	31
3.1 Considerações sobre o modelo de negócios.....	31
3.2 Panorama do mercado brasileiro de soluções para preparação ao vestibular baseadas em tecnologia	35
3.3 Panorama de mercado de soluções educacionais baseadas em LLMs	40
3.3.1 A hype dos Large Language Models.....	40
3.3.2 Panorama de mercado de soluções educacionais baseadas em LLMs.....	43
3.4 Análise estratégica	46
4 MVP: ESTRUTURAÇÃO E PLANEJAMENTO	53
4.1 Entendendo o papel de um MVP	53

4.2	Formulação da tese e hipóteses	54
4.3	Definição de experimentos	56
4.3.1	Hipótese 1 (técnica).....	56
4.3.2	Hipótese 2 (demanda).....	61
4.3.3	Hipótese 3 (efetividade)	62
4.3.4	Hipótese 4 (viabilidade econômica).....	64
4.4	Especificação do MVP e orçamento inicial.....	64
4.4.1	Visão geral	64
4.4.2	Base de Dados e Dataset "ENEM Challenge"	65
4.4.3	Sistema Recomendador de Questões.....	65
4.4.3.1	Sistema Recomendador Baseado em TRI	65
4.4.3.2	Sistema Recomendador Alternativo	Erro! Indicador não definido.
4.4.4	Integração com GPT-4 para Geração de Respostas.....	67
4.4.5.1	Formulários de Personalização	67
4.4.6	Desenvolvimento Ágil e iterativo	68
4.4.6.1	Estrutura do Ciclo de Sprints	68
4.4.7	Infraestrutura Tecnológica	69
4.4.8	Orcamento e Custos Estimados.....	69
4.4.8.1	Custo de Desenvolvimento.....	69
4.4.8.2	Infraestrutura Tecnológica e Ferramentas Pagas	69
4.4.8.3	Budget de Divulgação.....	70
4.4.9	Indicação de um <i>roadmap</i>	70
CONCLUSÃO.....		72
REFERÊNCIAS		74
ANEXO A – Edital Itaú IA para a Educação (2024).....		78

1 INTRODUÇÃO

1.1 O sentido do uso de tecnologia no combate à desigualdade de oportunidades

A desigualdade de oportunidades no Brasil é uma herança resiliente do passado colonial e escravocrata do país, profundamente enraizada nas suas estruturas econômicas e sociais, que continua a minar as perspectivas de amadurecimento da democracia brasileira. Existem inúmeras pesquisas teóricas e empíricas que apontam que fatores além do controle individual, especialmente a posição socioeconômica ao nascer, têm um impacto profundo nas perspectivas de vida dos brasileiros. Um estudo da OCDE de 2018 sobre mobilidade social revelou que, em média, são necessárias nove gerações para que os 10% mais pobres alcancem a classe média. Esse dado expõe uma realidade em que a posição social de milhões de pessoas é predeterminada por circunstâncias alheias à sua conduta, uma grave erosão da agência individual que se choca com o princípio de liberdade e a concepção de direitos humanos. Ademais, hoje é praticamente um consenso entre cientistas sociais e economistas que a perpetuação da desigualdade social gera consequências devastadoras para a coesão social e o desenvolvimento econômico, prejudicando a sociedade como um todo e sendo uma questão de interesse público.

O enfrentamento dessa desigualdade e de seus efeitos nocivos certamente requer reformas estruturais profundas, mas, em um cenário de recursos limitados, a adoção estratégica de tecnologias emergentes pode amplificar significativamente os esforços de transformação social. Desde a década de 1960, os avanços em tecnologias computacionais têm revolucionado a produção, o armazenamento e a transmissão de informações, transformando setores inteiros da economia. Este trabalho parte do entendimento de que existe um grande potencial inexplorado na aplicação dessas tecnologias pelo setor público com a finalidade de contribuir com a promoção da justiça e a transformação da sociedade brasileira. Não se trata de defender que o estado deva ser buscar uma participação direta no desenvolvimento de novas tecnologias e na inovação, mas sim utilizar as vantagens que elas podem oferecer no contexto público, aproveitando ensinamentos de aplicações comprovadas no setor privado e na experiência internacional em prol do enfrentamento de desafios coletivos. Em particular, podem ser destacadas as aplicações no setor privado que utilizam o poder de escala e adaptabilidade das soluções digitais no contexto educacional, oferecendo uma inspiração clara para seu uso estratégico em políticas públicas.

Ao considerar o problema da desigualdade de oportunidades no Brasil, é natural direcionar o foco para o sistema educacional do país. A experiência internacional ao longo do século XX

demonstrou repetidamente que investimentos estratégicos em educação pública não só são essenciais para a redução das desigualdades sociais, mas também constituem a base para o desenvolvimento econômico sustentável e a prosperidade de nações. Países que priorizaram a educação em suas políticas públicas, como Finlândia e Coreia do Sul, colheram benefícios significativos tanto em termos de coesão social quanto de crescimento econômico. No Brasil não é diferente, e a educação desempenha um papel crucial na perpetuação ou mitigação das desigualdades, considerando sua função natural de grande equalizador social. No seu estado atual, o sistema brasileiro ainda está parcialmente inserido na estrutura social que perpetua as desigualdades, mas esforços muito significativos tem sido feitos para mudar essa conjuntura, e os avanços apenas não são maiores pela falta de alinhamento político e a consequente limitação de recursos destinados a seu desenvolvimento.

1.2 A inserção do sistema educacional brasileiro nas estruturas que perpetuam desigualdades

A educação é fundamental para a integração dos indivíduos à sociedade, tornando-os aptos a contribuir para a produção econômica e a usufruir dela. Neste contexto, uma função essencial do sistema educacional é a de preparar os jovens, independentemente de suas origens, com conhecimentos práticos e teóricos necessários para um estabelecimento profissional futuro que atenda simultaneamente aos interesses individuais e sociais. Objetivando a justiça social, as oportunidades de trabalho deveriam ser determinadas predominantemente pelo desenvolvimento de cada indivíduo em suas aptidões e iniciativas primeiro durante o processo de formação, e depois na prática profissional. Contudo, a realidade é que a maioria dos jovens brasileiros depende de um ensino público notoriamente ineficaz, e fatores socioeconômicos, fora de seu controle, têm um papel determinante em suas perspectivas futuras.

O sistema educacional brasileiro é composto por duas grandes etapas: o ensino básico, que cobre conhecimentos fundamentais e gerais, e o ensino superior, que aprofunda conhecimentos em áreas específicas que possuem valor na produção econômica e cultural. A transição entre essas etapas, crucial para a qualificação do capital humano e, consequentemente, para a divisão de oportunidades econômicas, é mediada pelo exame vestibular. Esse exame, que busca avaliar objetivamente o aproveitamento do ensino básico pelos estudantes, desempenha um papel decisivo na trajetória educacional e profissional dos indivíduos. Para que seja um mecanismo

justo, as condições deveriam ser niveladas entre os candidatos, mas isso é muito distante do que ocorre na prática.

Isolando o ensino básico em nosso recorte, ele deveria em primeiro lugar contemplar todos os jovens brasileiros, e em segundo oferecer condições adequadas para o ensino superior, incluindo a preparação para o vestibular. Em relação à primeira condição, a lei brasileira impõe educação obrigatória e gratuita para todos os cidadãos entre 4 e 17 anos, abrangendo a educação básica, sendo a efetivação dessa lei uma conquista progressiva desde a Constituição de 1988. Tendo em vista a segunda condição, porém, a universalização da educação básica esbarra na questão da qualidade do ensino, que permanece um problema crônico. Avaliações internas, como dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2019)¹, indicam que 6 em cada 10 estudantes saem do ensino médio com conhecimentos insuficientes em matemática e português. No PISA (2023), uma avaliação externa, o Brasil apenas escapou do quartil inferior em leitura, entre 81 países avaliados. O Censo Escolar de 2023 mostra que 81,1% das matrículas no ensino básico são atendidas pela rede pública, enquanto o restante está na rede privada, predominantemente frequentada por alunos das classes A e B, que buscam maior qualidade de ensino e melhores oportunidades futuras. Em comparação, apenas 2 em cada 10 estudantes de escolas privadas apresentam conhecimentos insuficientes em português e matemática.

No ensino superior, por sua vez, não há obrigatoriedade, e o investimento estatal não se dilui em uma oferta tão ampla de vagas. As instituições públicas, que são as mais bem financiadas, formam os principais especialistas e produzem conhecimento relevante, e, por serem gratuitas, são naturalmente as mais concorridas. Nesse contexto, o vestibular se torna um mecanismo que discrimina com base em condições socioeconômicas, favorecendo jovens de origem privilegiada com acesso a um ensino básico privado, justamente os que teriam condições de pagar por educação superior privada, enquanto exclui a maior parte dos estudantes da rede pública e suas chances de ascensão social através dos estudos.

Cumprir os objetivos estabelecidos na Constituição Federal, que incluem a busca por uma educação efetiva e empoderadora e a oferta de oportunidades de qualificação para todos, depende da melhora do ensino básico e da ampliação das vagas no ensino superior. Isso, por sua vez, requer um aumento substancial do investimento público em educação e a implementação contínua de políticas educacionais ao longo de vários ciclos eleitorais. Esse tipo

¹ Relatório escolhido por ser relativo à última avaliação anterior à pandemia de COVID-19, cujos efeitos sobre aprendizado preferimos não considerar aqui

de mudança é lenta por natureza e depende de um alinhamento político que priorize a educação e a redução das desigualdades; é uma luta essencial de todos que buscam uma sociedade mais justa. No entanto, a reforma estrutural não é a única solução disponível. Com uma visão sistêmica, é possível identificar gargalos e implementar políticas de alto custo-benefício, que impactem a vida de milhares de jovens brasileiros que necessitam de mudanças imediatas, o que é comprovado pela história recente.

1.3 Democratização no acesso ao ensino superior como alvo de políticas públicas de alto custo benefício

Como apresentado, um dos principais gargalos do sistema educacional brasileiro segundo sua própria estrutura é o acesso ao ensino superior por estudantes oriundos do ensino público básico. Felizmente, ao longo da última década, políticas de alto custo-benefício direcionadas a esses estudantes mostraram sua efetividade, servindo como exemplos valiosos para nós.

Atualmente, existem três grandes programas governamentais voltados para o acesso ao ensino superior, todos utilizando o Exame Nacional do Ensino Médio (ENEM) como mecanismo de seleção. Dois desses programas se concentram nas universidades privadas: o Fundo de Financiamento ao Estudante do Ensino Superior (FIES), criado em 1999, que oferece financiamento total ou parcial das mensalidades, e o Programa Universidade para Todos (PROUNI), de 2004, que oferece bolsas parciais ou integrais para estudantes com renda familiar per capita de até três salários mínimos. O terceiro programa, o Sistema de Seleção Unificada (SiSU), criado em 2010, facilita o acesso a universidades públicas em todo o país através de um processo seletivo unificado. Contudo, esse progresso em democratizar o acesso ao ensino superior só se tornou significativo com a implementação da Lei de Cotas.

Instituída em 2012, a Lei nº 12.711, popularmente conhecida como Lei de Cotas, determinou que 50% das vagas em instituições federais de ensino técnico e superior fossem reservadas a estudantes de escolas públicas. Essa lei também influenciou instituições estaduais a adotarem políticas semelhantes nos anos subsequentes (TOLEDO, 2024). Na época de sua implementação, houve questionamentos sobre a capacidade acadêmica desses estudantes em universidades de excelência. No entanto, mais de 10 anos depois, os dados são claros em demonstrar que esses estudantes têm grande capacidade de superar as dificuldades iniciais, frequentemente apresentando desempenho superior ao de colegas de origem mais privilegiada (MUGNAINI JUNIOR, 2023).

Os programas mencionados, ao reservarem metade das vagas em universidades públicas para estudantes de escolas públicas e ao promoverem o financiamento para cursos privados, ampliaram as oportunidades de qualificação de capital humano para brasileiros antes excluídos dessas possibilidades, efetivamente contribuindo para a redução da desigualdade de oportunidades e utilizando os recursos de maneira eficiente. Embora os efeitos na mobilidade social ainda estejam em processo de avaliação, tudo indica que a transição do ensino básico para o ensino superior é um ponto crítico para intervenções voltadas à democratização das oportunidades, e ainda há muito a ser feito nessa direção.

Apesar da criação de canais de acesso à universidade para os $\frac{4}{5}$ dos jovens brasileiros que frequentam o ensino básico público, é crucial considerar os recursos disponíveis a estes estudantes para que eles possam competir pelas vagas disponíveis, cujo crescimento é limitado pelo investimento público. Esse vasto grupo de alunos é marcado por grandes desigualdades socioeconômicas e regionais, considerando as diferenças entre as redes municipal, estadual e federal (esta última responsável por uma pequena fração do ensino médio) que compõem o sistema público de ensino básico. Todos esses estudantes enfrentam os mesmos exames vestibulares, com destaque para o ENEM, que funciona como um filtro determinante para o acesso à universidade, e a falta de recursos acessíveis de preparação fora da escola representa um problema significativo de agência que afeta aqueles com pouca disponibilidade de recursos.

Em um passado recente, poderia parecer irracional direcionar dinheiro público que poderia ser empregado na melhoria da educação básica para a preparação para o vestibular. Contudo, com as tecnologias digitais atuais, é possível oferecer recursos instrucionais em grande escala, amplamente utilizados pelo setor privado em produtos voltados para essa finalidade. A utilização dessas tecnologias pelo poder público pode proporcionar aos estudantes motivados as condições necessárias para reduzir a distância entre seus sonhos e a realidade, demonstrando-se uma política educacional com um custo-benefício social altamente interessante.

1.4 Desafios à agência dos estudantes na preparação para o vestibular e o potencial de soluções a partir de tecnologia

Apesar dos avanços nas políticas que visam ampliar o acesso ao ensino superior para estudantes oriundos de escolas públicas, é crucial reconhecer que ainda há um longo caminho a ser percorrido para garantir sua plena efetividade. Um dos maiores desafios é a permanência desses estudantes na universidade, especialmente daqueles que não possuem renda suficiente

para se sustentar. A realidade financeira de muitas famílias brasileiras obriga seus filhos a trabalhar assim que possível para contribuir com a renda familiar, o que é uma das principais causas de evasão no ensino médio. Quando esses jovens superam essa barreira e conseguem uma vaga na universidade, muitas vezes em cidades distantes, enfrentam o desafio de se manterem longe de casa. A maioria das universidades públicas não oferece moradias estudantis ou suporte financeiro suficientes para atender a todos os estudantes que necessitam, e os programas de estágio disponíveis raramente são compatíveis com as demandas acadêmicas e as necessidades financeiras desses alunos. Além disso, falta incentivo no ensino público para que os jovens se preparem adequadamente para os vestibulares e aspirem ao ensino superior, o que compromete a inclusão e a democratização do acesso à educação de qualidade.

Do ponto de vista individual, a situação é tão complexa e problemática que o benefício de incentivar o ingresso no ensino superior nem sempre é claro. Embora seja comprovado que a educação, especialmente o ensino superior e a especialização profissional, é uma das principais ferramentas de ascensão social, essa consciência pode se tornar uma fonte de frustração diante das inúmeras barreiras existentes. Considere o caso de João, um estudante hipotético do último ano do ensino médio, que apesar de almejar a universidade, precisa dividir seu tempo entre trabalho e escola, uma instituição pública que não oferece preparação específica para o vestibular. Considerando os recursos disponíveis a ele, João larga muito atrás na competição pelas vagas no ensino superior.

Em contraste, jovens de famílias mais privilegiadas no contexto social brasileiro, além de não precisarem trabalhar, podem dedicar seu tempo integralmente aos estudos em escolas privadas que, em geral, oferecem uma preparação para o vestibular com resultados demonstrados. Esses estudantes também têm acesso a uma ampla gama de soluções de preparação disponíveis no mercado, como os cursinhos pré-vestibulares, uma indústria consolidada e tradicional, que ao longo dos anos aperfeiçoou seus métodos por meio da experiência, oferecendo uma combinação eficaz de aulas presenciais e material didático direcionado exclusivamente para os exames. Nos últimos 15 anos, essas opções foram complementadas por soluções inovadoras baseadas em plataformas digitais, que adaptam o modelo comprovado dos cursinhos à escalabilidade da tecnologia, proporcionando acesso mais amplo, embora ainda limitado a quem dispõe de recursos financeiros e tempo para se beneficiar plenamente.

Há muitas pessoas e organizações mobilizadas para viabilizar alternativas para casos como o de João, que representa milhares de jovens espalhados pelo Brasil. Essas iniciativas incluem organizações do terceiro setor, projetos sociais oriundos de comunidades ou das próprias

universidades, como os cursinhos populares, além de fundações filantrópicas. No entanto, a maioria dessas opções é restrita ao ambiente presencial, o que limita sua escala e acessibilidade devido a barreiras físicas. É nesse contexto que faz sentido a proposta de agregar recursos públicos e utilizar o alcance universal da tecnologia para viabilizar recursos gratuitos de preparação para o vestibular, acessíveis a todos os estudantes em qualquer lugar e a qualquer momento, desde que disponham de um dispositivo computacional, como um celular, e uma conexão com a internet. É amplamente reconhecido que essa restrição tecnológica é significativa, podendo impedir que uma parcela dos jovens acesse a solução. No entanto, essa limitação deve ser abordada por outras políticas públicas focadas na inclusão digital, que é um elemento cada vez mais essencial no mundo contemporâneo.

Há diversas maneiras com que recursos públicos podem ser utilizados para viabilizar tecnologias voltadas à preparação para o vestibular. Ao lidar com tecnologia, o Estado deve aproveitar a experiência acumulada pela iniciativa privada para mitigar riscos e maximizar a eficácia das políticas, e nesse sentido, pode ser extremamente vantajoso replicar modelos bem-sucedidos, como o dos cursinhos EAD. No entanto, o foco deste trabalho recai sobre o potencial das novas tecnologias baseadas em Inteligência Artificial (IA), especialmente a IA Generativa, que não só permite a disponibilização de recursos em larga escala e com grande flexibilidade, superando barreiras físicas, mas também oferece a possibilidade inédita de assistência individualizada. As características dessas tecnologias são particularmente interessantes para lidar com provas padronizadas, como as do vestibular, e por isso os esforços aqui empreendidos foram focalizados em desenhar uma solução viável do ponto de vista técnico e um modelo de execução adequado à realidade brasileira.

1.5 Proposta de uma solução de tutoria para o ENEM baseada em Large Language Models

Diante do desafio de combater a desigualdade de oportunidades por meio da democratização dos recursos de preparação para o vestibular, é estratégico direcionar nossos esforços ao ENEM, a principal porta de entrada para as universidades públicas brasileiras. Em seu formato atual, o exame consiste em duas provas com 90 questões de múltipla escolha - uma abrangendo linguagens e ciências humanas, e outra focada em matemática e ciências da natureza - além de uma prova de redação. Esse modelo foi projetado para facilitar a correção em larga escala, mas

também oferece características extremamente para o uso das capacidades inovadoras da Inteligência Artificial Generativa, especialmente através dos Large Language Models (LLMs).

No âmbito da IA Generativa, que compreende os sistemas baseados em Inteligência Artificial capazes de produzir novos conteúdos digitais, como textos, imagens, áudios e vídeos, os LLMs destacam-se por sua impressionante capacidade de interpretar e gerar linguagem de forma coerente. Esses modelos estão revolucionando o campo da Inteligência Artificial ao simular a cognição humana e aumentar muito as possibilidades de interação entre computadores e usuários, inclusive por meio de conversação direta. Em contextos educacionais, pela primeira vez há máquinas capazes de estabelecer diálogos com os estudantes, mediando a relação com todo o conhecimento de livros, artigos e até vídeos armazenados digitalmente. Ao viabilizar a reprodução de características essenciais da tutoria (um método educacional baseado na atenção individual e comprovadamente eficaz) em escala, os LLMs se tornam uma ferramenta extremamente poderosa para apoiar o aprendizado e democratizar o acesso ao conhecimento (KHAN, 2024).

Os Large Language Models do estado da arte atual possuem um vasto repertório de conhecimentos adquiridos em seu treinamento, e são comumente chamados Modelos Fundacionais por servirem como base para diversas aplicações. A flexibilidade característica a esses modelos permite que interpretem corretamente e resolvam questões relacionadas a conhecimentos básicos, como as do ENEM; evidências dessa competência serão apresentadas em nosso referencial teórico. As próprias informações contidas nos enunciados e nas alternativas (considerando o gabarito) ajudam os modelos a se situar no domínio apropriado de conhecimento. Além de resolver as questões, conseguem abordá-las por diferentes perspectivas, permitindo que respondam a dúvidas individuais e únicas, algo que materiais didáticos estáticos não podem oferecer. A essa flexibilidade, corresponde uma certa inconsistência: não raramente, LLMs reproduzem informações falsas com extrema confiança, fenômeno conhecido como “alucinações”. Esse problema será discutido com indicações de algumas técnicas com potencial de mitigá-lo.

Provas padronizadas como o ENEM têm como objetivo avaliar o conhecimento dos estudantes, mas muitas vezes, as habilidades para lidar com o formato da prova se revelam tão importantes quanto o conteúdo. Por conta disso e pelas próprias características dos LLMs, nossa proposta de solução será focada em uma prática assistida para a prova. Além de oferecer orientações sobre o conteúdo, os LLMs podem ser programados para fornecer dicas de resolução, ajudando os estudantes a desenvolver estratégias eficazes para enfrentar o formato

específico das questões do ENEM. Vemos o formato de Chatbot, com que o público já está amplamente familiarizado, como a melhor opção de interface para sustentar essa prática.

O uso de sistemas computacionais nesse formato possui outra vantagem significativa: a possibilidade de personalização da prática via recomendação de questões baseada em dados. A abordagem tradicional na preparação para o vestibular consiste na imposição dos mesmos conteúdos teóricos e exercícios práticos para todos os estudantes, o que leva a um subaproveitamento. Em ambientes digitais, é possível capturar e processar os dados das interações individuais através de vários métodos para recomendar conteúdos específicos segundo critérios que visam otimizar o aproveitamento. Entre os métodos existentes, destacam-se os baseados em Teoria de Resposta ao Item (TRI), que já é a base para a avaliação das provas do ENEM, e permite levar em conta a dificuldade das questões e o nível de conhecimento atual de cada estudante (LORD, 1980). Idealmente, a expectativa é que o sistema possa recomendar as questões que têm maior potencial de melhorar o desempenho de cada usuário em dado momento. Aliando isso à tutoria dedicada de um sistema de IA que oferece suporte contínuo e ajustado às necessidades específicas de cada estudante, acredita-se ser possível desenvolver uma solução voltada a preparação para o ENEM muito direcionada e eficaz, disponível a todos os estudantes da rede pública.

A solução proposta é composta por três elementos básicos principais: uma base de dados com as questões históricas do ENEM, um sistema recomendador de questões e a integração com um LLM do estado da arte em uma estrutura de Chatbot. Em sua versão mais simples, será um sistema que estudantes poderão acessar a qualquer momento por meio de celular ou computador conectado à internet, e interagir com um tutor baseado em Inteligência Artificial para praticar com questões do ENEM e ganhar familiaridade progressiva com a prova. Tanto o sistema recomendador quanto a integração com o LLM poderão ser aprimoradas gradativamente, sendo fundamental validar sua entrega de valor para estudantes ao longo das etapas dessa progressão, algo que trataremos extensivamente. Uma solução final poderá estimar o desempenho do usuário na prova a partir de uma amostra reduzida de questões, recomendar questões que melhor se adequem ao nível de preparo de cada estudante, e utilizar a Inteligência Artificial de ponta para fornecer orientação detalhada e personalizada em cada questão, atendendo a dúvidas específicas e ajudando o estudante a desenvolver tanto o conhecimento quanto as habilidades necessárias para enfrentar o exame. Ela se enquadra num modelo de ensino adaptativo e personalizado, fundamentado por um longo histórico de pesquisa e desenvolvimento de soluções educacionais, e poderá ser oferecida em ampla escala, com um custo reduzido por usuário.

Embora a solução em seu formato final tenha um alto grau de complexidade, sua versão inicial pode ser bastante simples em relação aos benefícios potenciais, e enfatizamos a estratégia de um desenvolvimento progressivo voltado à validação de hipóteses na exposição com usuários reais, assegurando o uso eficiente dos recursos e mitigando riscos ao longo do desenvolvimento. O objetivo geral desse trabalho é o de estabelecer as bases para a viabilização da solução descrita, com os objetivos específicos de (1) justificar sua pertinência perante a sociedade e como alvo de financiamento público, (2) apresentar seus principais fundamentos teóricos por meio de extensa pesquisa, (3) propor um modelo de negócios viável ancorado em experiências da iniciativa privada e (4) validar algumas das principais hipóteses e propor um planejamento para seu desenvolvimento. A longo prazo, caso esse projeto possa ser realizado, espera-se que esta solução possa empoderar estudantes com menos oportunidades, oferecendo-lhes condições para uma preparação efetiva ao vestibular e consequentemente para o acesso ao ensino superior, contribuindo assim para uma sociedade brasileira mais acolhedora e democrática.

Na seção 2, serão abordados os principais referenciais teóricos relacionados ao projeto. Primeiro, serão discutidas teorias educacionais atreladas à computação, que servem de fundamento para a solução descrita. Em seguida, serão explorados alguns dos aspectos centrais sobre LLMs que fazem deles úteis em nossa aplicação, e técnicas relevantes para a adaptação de modelos generalistas para suas especificidades. Por fim, serão apresentados estudos que testaram modelos de última geração em questões do ENEM, mostrando seus resultados.

Na seção 3, serão abordadas possibilidades para o modelo de negócios que poderá tornar a solução viável. Em seguida, o foco será colocado em uma pesquisa de mercado, primeiro com o mapeamento de algumas das principais soluções baseadas em tecnologia para a preparação de vestibulares, depois com a apresentação do cenário de LLMs e soluções educacionais promissoras, sobretudo voltadas a tutoria, baseadas em nesse tipo de IA. Essa parte será finalizada com uma análise estratégica : sobre o posicionamento do projeto frente a alternativas, suas forças e fraquezas, e o detalhamento do modelo proposto tendo em vista relações com parceiros e clientes, atividades, e também considerando o aspecto financeiro.

Na seção 4, será introduzido o método *Lean Startup* para estabelecer os fundamentos de um modelo de desenvolvimento voltado à mitigação de riscos. As hipóteses centrais da tese serão elaboradas e experimentos propostos para validá-las. Em seguida, será especificado o formato inicial para a solução, ou MVP, e introduzido um planejamento para sua execução, contemplando as principais atividades e um orçamento dos recursos financeiros necessários.

2 REFERENCIAL TEÓRICO

Nesta seção, cobriremos alguns dos principais aspectos teóricos que se relacionam com nossa solução. Primeiro, abordaremos brevemente algumas concepções conceituais que historicamente foram relevantes para direcionar a aplicação de tecnologia computacional para fins educacionais, bem como seu contexto. Em seguida, nos aprofundaremos nos LLMs, nas propriedades que fazem desse tipo de sistema muito adequado para fins de tutoria, e algumas das técnicas utilizadas para adaptar seu comportamento. Por fim, buscaremos na literatura elementos que possam validar nossa hipótese da utilização de modelos fundacionais para criar explicações com relação ao ENEM.

2.1 Teorias para o uso de tecnologia computacional na educação

Há uma vasta e contínua experiência teórica e prática no uso de tecnologias da informação em contextos educacionais, refletindo a evolução significativa dessas tecnologias ao longo das últimas décadas. O período que seguiu o término da Segunda Guerra Mundial foi marcado por inovações tecnológicas revolucionárias, como a invenção dos transistores, o desenvolvimento de circuitos integrados, a criação dos microprocessadores e a introdução das interfaces gráficas. Essas inovações transformaram os computadores em Tecnologias de Propósito Geral (General Purpose Technologies), tornando-os programáveis, acessíveis e amigáveis para usuários leigos.

Essas máquinas, capazes de armazenar, processar e transmitir informações de diversas maneiras, rapidamente despertaram o interesse de educadores e pesquisadores, que reconheceram seu imenso potencial para a aplicação em atividades educacionais. Desde então, tem havido um percurso de experimentação e desenvolvimento, com o acúmulo de uma rica experiência no uso da tecnologia computacional para facilitar e aprimorar os processos de ensino e aprendizagem. A educação assistida por computador passou por várias fases de evolução, desde as primeiras máquinas de ensino até os sofisticados ambientes de aprendizagem online de hoje, refletindo um progresso contínuo e uma adaptação às necessidades pedagógicas em constante mudança.

A ideia de que computadores poderiam ser utilizados para otimizar aspectos centrais da aprendizagem individual, bem como aumentar a eficiência da educação no contexto social remontam a artigos escritos na década de 1950 por B. F. Skinner. O psicólogo pai do behaviorismo descreveu “máquinas de ensinar” que permitiriam que estudantes recebessem

feedbacks instantâneos relacionados ao seu conhecimento em vez do retorno dilatado das tradicionais lições de casa e exames. Essa tecnologia incorporaria os princípios das teorias de aprendizagem por reforço de Skinner, que estabelecem a importância de estímulo reforçado para constituir um comportamento desejado, aqui aplicado em relação a aprendizagem correta de conceitos e técnicas. Ademais, máquinas de tal tipo seriam capazes de possibilitar um ritmo de progressão adequado às características individuais de cada estudante, ao invés de impor um ritmo comum (próprio à sala de aula), que seria notoriamente contraproducente para o autor.

No trecho a seguir, Skinner compara as propriedades de tal máquina com as de um (bom) tutor:

“The comparison holds in several respects. (1) There is a constant interchange between program and student. Unlike lectures, textbooks, and the usual audiovisual aids, the machine induces sustained activity. The student is always alert and busy. (2) Like a good tutor, the machine insists that a given point be thoroughly understood, either frame by frame or set by set, before the student moves on. Lectures, textbooks, and their mechanized equivalents, on the other hand, proceed without making sure that the student understands and easily leave him behind. (3) Like a good tutor the machine presents just that material for which the student is ready. It asks him to take only that step which he is at the moment best equipped and most likely to take. (4) Like a skillful tutor the machine helps the student to come up with the right answer. It does this in part through the orderly construction of the program and in part with such techniques as hinting, prompting, and suggesting, derived from an analysis of verbal behavior (47). (5) Lastly, of course, the machine, like the private tutor, reinforces the student for every correct response, using this immediate feedback not only to shape his behavior most efficiently but to maintain it in strength in a manner which the layman would describe as “holding the student’s interest.”” (Skinner, 1958)

No mesmo artigo, Skinner destacava que essa tecnologia educacional deveria ser vista como um complemento ao papel do educador, e não como um substituto. Ele argumentava que seu uso poderia melhorar a aprendizagem dos estudantes de maneiras que os professores, por si só, não poderiam alcançar, ao mesmo tempo em que liberaria os professores para se dedicarem a contribuições que só eles poderiam fazer. O aspecto da escala também era central na visão do autor: ao contrário dos materiais físicos, como livros e apostilas, programas de computador e dados podem ser replicados a um custo marginal próximo de zero. Esses princípios de economias de escala foram e continuam sendo um elemento fundamental na história da aplicação de tecnologias computacionais à educação, ou talvez principalmente, nas esperanças depositadas em seu potencial. Como é comum na história da computação, as dificuldades de explorar todo potencial envisionado se mostraram muito maiores do que Skinner previa; a base tecnológica ainda era insuficiente. Mas isso não significou uma perda de

força das suas ideias, que seguiram inspirando reflexões sobre aplicação de tecnologia na educação.

30 anos depois, os princípios estabelecidos por Skinner sobre a efetividade da aprendizagem através do reforço podem ser relacionados com estudos sobre abordagens educacionais como o *mastery learning* e a tutoria individualizada. A prática de *mastery learning* baseia-se na premissa de que cada aluno deve dominar completamente um conceito antes de avançar, e a de tutoria individualizada agrega a isso o fornecimento feedback imediato e ritmo de aprendizagem adaptado às características individuais do estudante. Na década de 1980 foram realizados estudos que demonstraram que essas abordagens resultam em ganhos significativos no desempenho acadêmico dos estudantes, de respectivamente 1 e 2 desvios padrões acima daqueles que seguem o ensino tradicional em sala de aula (ANANIA, 1982,1983; BURKE,1984).

O pedagogo Benjamin Bloom, ao analisar os resultados positivos do *mastery learning* e da tutoria individualizada, destacou a importância social de tornar esses métodos altamente eficazes acessíveis a um número maior de estudantes. Em artigo de 1984, Bloom argumentou que, se fosse possível escalar essas abordagens para o ensino em grupo, seria possível elevar significativamente o desempenho acadêmico de grandes populações, aumentando a produtividade econômica, reduzindo desigualdades educacionais e proporcionando oportunidades mais equitativas. No entanto, ele também enfatizou o grande desafio desse empreendimento: a falta de tecnologias e recursos humanos capazes de replicar o impacto profundo da tutoria individualizada em larga escala. Bloom apresentou o "problema dos 2 sigmas", que ficou conhecido como *Bloom-2-Sigma problem* como uma questão central na busca por métodos de ensino que possam combinar eficácia e acessibilidade. Desde então, sua análise influenciou inúmeras pesquisas e desenvolvimentos em tecnologias educacionais, que continuam a explorar formas de democratizar o acesso à educação de alta qualidade, na esperança de alcançar os benefícios que Bloom vislumbrou para a sociedade (Bloom, 1984).

Desde então, mais uma série de inovações se acumularam e contribuíram para que as tecnologias computacionais façam mais parte da vida das pessoas que nunca. O aumento exponencial da capacidade de armazenamento e processamento dos computadores abriu novas possibilidades para o desenvolvimento de software educacional mais sofisticado e adaptativo. A internet, junto com o aumento na velocidade de transmissão de dados, rompeu barreiras geográficas, permitindo a produção e o compartilhamento coletivo de conhecimento em uma escala global sem precedentes. Além disso, a popularização dos smartphones democratizou o acesso à tecnologia computacional e à internet, colocando ferramentas de aprendizado

avançadas nas mãos de pessoas em todo o mundo, independentemente de sua localização ou condição econômica. Na década de 2010 ocorreu uma explosão de empresas de tecnologia educacional, as chamadas EdTechs, com diferentes modelos de aprendizagem.

Mas havia um elemento que faltava: computadores seguiam sendo extremamente limitados na compreensão de linguagem natural e na sua reprodução. Sendo a língua o principal meio de interação do ser humano, fica muito claro que a distância de tecnologias computacionais em relação a práticas educacionais desenvolvidas por professores não se limitava à efetividade humana. Havia um obstáculo técnico extremamente relevante, que acaba com a chegada dos LLMs; estes abrem todo um novo horizonte para a exploração de sistemas automatizados de ensino.

2.2 Especificando LLMs e principais técnicas para sua adaptação

2.2.1 Sobre a origem e principais propriedades dos LLMs

Os Large Language Models (LLMs) surgiram como uma inovação significativa no campo da inteligência artificial, resultado direto dos avanços em modelos de linguagem baseados em uma arquitetura denominada *Transformers*. A introdução do modelo de Transformer por Vaswani et al. em 2017, com seu mecanismo de atenção, revolucionou a forma como as redes neurais processam e geram linguagem natural, permitindo a construção de modelos que podem ser treinados com efetividade em vastas quantidades de dados textuais. Esses modelos, ao aprenderem padrões complexos de linguagem, evoluíram rapidamente em capacidade e sofisticação. Os LLMs tornaram-se amplamente conhecidos através da série de modelos GPT (*Generative Pre-trained Transformer*), desenvolvidos pela OpenAI, sendo o GPT-3 (Brown et al., 2020) e, posteriormente, o ChatGPT (2022) os marcos que popularizaram esses modelos ao público global.

LLMs pertencem à classe dos Modelos Neurais de Linguagem (*Neural Language Models*), por sua vez inseridos no paradigma mais amplo da modelagem de linguagem (*language modeling*), que compreende décadas de esforços de pesquisadores de modelar a linguagem natural (língua) através de modelos matemáticos e estatísticos. Enquanto Modelos Neurais, LLMs são essencialmente redes neurais profundas que recebem como input quantidades delimitadas de texto e devem produzir como output a continuação do texto, utilizando para isso

seus parâmetros, pesos e viés que conectam os neurônios artificiais entre suas camadas e que são aprendidos na fase de treinamento da rede. Os LLM recebem esse nome porque seu grande tamanho é extremamente relevante: possuem “a partir de dezenas de bilhões ou centenas de bilhões de parâmetros”, e a literatura identifica que seu tamanho se relaciona diretamente com as propriedades que fazem deles únicos (Zhao et al., 2023).

Aqui não é o caso de nos aprofundarmos na arquitetura dos LLMs, explicar o que são os *Transformers*, ou como palavras são transformadas em números comprehensíveis pelos modelos. É de nosso interesse o entendimento das possibilidades de aplicação desse tipo de sistema de IA, quais são suas capacidades e limitações. O que torna os LLMs tão especiais é que a partir de uma certa escala de Modelos Neurais de Linguagem - volume de dados, quantidade de computação empregada, número de parâmetros do modelo - estes começam a apresentar capacidade de produzir texto de maneira praticamente indistinguível de seres humanos. Até a emergência desta tecnologia, a produção de texto era uma das tarefas tidas como mais representativas da cognição humana, pois existem uma série de operações cognitivas subjacentes a ela: operações lógicas, de articulação de sentido, emprego de estilo, etc. Mas a partir do momento que redes neurais se tornaram aptas a aprender consistentemente padrões da forma textual, verificou-se que se tornam capazes de manipular textos mantendo a coerência lógica, de sentido e estilo a partir dos exemplos a que foram expostas; são essas habilidades que os tornam potencialmente muito úteis para ajudar estudantes a compreender individualmente conteúdos subjacentes a questões como as do ENEM.

Essas habilidades são adquiridas na fase chamada “pré-treinamento”, quando as gigantescas redes neurais (baseadas nos *Transformers*) são treinadas a partir de extensiva computação (para treinar modelos de ponta, são necessários centenas de milhões de dólares apenas em gastos computacionais) em córpus (conjuntos de dados textuais) tão vastos que considera-se que compreendem praticamente todos os textos contidos na internet. Para nossos fins, é essencial que o modelo tenha sido exposto no pré-treinamento aos conhecimentos que serão exigidos pelas questões, e que sua complexidade seja suficiente para ser efetivo em ajustar esses conhecimentos a diferentes contextos. Para verificar isso, não há forma melhor que testar modelos nas provas, e apresentaremos estudos nesse sentido nas próximas seções.

Existem também outras capacidades típicas de modelos grandes (chamadas ‘habilidades emergentes’ pela literatura) que nos interessam bastante, como aprendizagem em contexto (*In-context learning*) e prompt com corrente-de-pensamento (*Chain-of-Thought prompting*, ou CoT). Mas antes, precisamos abordar as etapas posteriores de desenvolvimento de LLMs. Na fase de pré-treinamento os modelos aprendem a completar um pedaço de texto que lhes foi dado

como input a partir de pura probabilidade, mas para que eles sejam aptos a responder instruções em formato de pergunta-resposta e para que se comportem atendendo a outros requisitos desejáveis, como de segurança, passam respectivamente pelas etapas de ajuste fino (*fine-tuning*) e aprendizado de reforço a partir de feedback humano (*reinforcement learning through human feedback*, ou RLHF). Essas duas etapas compõem a fase de adaptação, em que os parâmetros do modelo passam por ajustes para que estes sejam capazes de aplicações gerais, como os Chatbots atuais. Após pré-treinamento e adaptação, temos um Modelo Fundacional, como GPT-3 ou GPT-4.

2.2.2 Alternativas para adaptação de LLMs a *use cases* específicos

Ao considerar a implementação de soluções baseadas em Large Language Models, é essencial entender que existem diferentes abordagens que impõem variados níveis de complexidade e custos ao projeto. As principais opções incluem o treinamento de um modelo do zero, o *fine-tuning* e o *reinforcement learning through human feedback* em modelos pré-treinados que permitam esses refinamentos, ou o uso de APIs que possibilitam a interação com modelos prontos através de *prompting*. Cada uma dessas abordagens oferece vantagens específicas, mas também apresenta desafios que precisam ser cuidadosamente considerados no contexto de um projeto como o nosso.

O treinamento de um modelo do zero é, sem dúvida, a abordagem mais complexa e custosa. Como foi mencionado, treinar um LLM de ponta pode envolver o uso de supercomputadores e centenas de milhões de dólares em recursos computacionais, e a quantidade de dados necessária para treinar esses modelos é imensa, exigindo um corpus que comprehenda vastas quantidades de texto para garantir que o modelo tenha uma base robusta de conhecimento. A complexidade desse processo é proibitiva para a maioria dos projetos, e é por isso que apenas algumas poucas empresas no mundo são capazes de realizar esse tipo de treinamento. Na verdade, o objetivo dessas empresas ao realizar tal treinamento é justamente criar Modelos Fundacionais que possam ser utilizados como base por uma miríade de empresas e organizações com aplicações direcionadas, como a nossa.

Em contraste, o *fine-tuning* e o RLHF oferecem formas já mais acessíveis de ajustar um modelo pré-treinado para atender às necessidades específicas de um projeto. O *fine-tuning* permite que o modelo seja ajustado para se comportar de uma maneira mais bem definida e alinhada ao tipo de interação que desejamos com os usuários. Por exemplo, podemos utilizar

dados de interações anteriores bem sucedidas ou dados coletados através de parcerias para ensinar o modelo a responder de uma forma mais adequada ao nosso contexto educacional. O RLHF, por outro lado, é um método que envolve a aplicação de feedback humano para refinar as respostas do modelo. Este processo pode ser muito valioso para garantir que o modelo apresente um comportamento adequado em diferentes contextos, por exemplo respondendo de maneira apropriada a diferentes tipos de perguntas relacionadas às questões do ENEM, mas depende de um esforço significativo em termos de coleta de feedback de qualidade e da participação de avaliadores humanos, o que aumenta o custo e a complexidade do projeto.

Modelos pré-treinados de ponta *open-source*, como os da família LLaMA (Large Language Model Meta AI), são candidatos promissores para serem submetidos a *fine-tuning* ou RLHF. Esses modelos já possuem capacidades avançadas devido ao extenso pré-treinamento realizado, mas seria necessário testá-los para verificar se o corpus utilizado no pré-treinamento os capacita a responder efetivamente às questões do ENEM. Além disso, essas abordagens são consideravelmente mais caras e técnicas, exigindo não apenas acesso ao modelo, mas também uma infraestrutura robusta e expertise em engenharia de *machine learning*. O *fine-tuning* e o RLHF também aumentam a complexidade do gerenciamento de dados, pois necessitam de conjuntos de dados adequados e processos contínuos de avaliação para assegurar a qualidade dos ajustes realizados.

Diante dessas considerações, uma abordagem inicial muito mais versátil e que oferece muitas possibilidades para validar nossas hipóteses centrais é o uso de APIs com *prompting*. As APIs fornecidas por provedores como OpenAI permitem o acesso direto a modelos altamente avançados sem a necessidade de gerenciar a infraestrutura complexa ou realizar ajustes finos. O *prompting* permite que personalizemos o comportamento do modelo de forma simplificada, dinâmica e interativa, ajustando os *prompts* para direcionar as respostas conforme necessário para o nosso projeto. Apesar de ter um custo variável maior, que pode ser desvantajoso no uso sustentado no tempo, o uso de APIs é muito adequado como ponto de partida, pois reduz significativamente os custos e a complexidade de desenvolvimento, enquanto ainda possibilita testar e validar os principais aspectos da nossa proposta. Por isso, o uso de APIs com modelos do estado da arte e técnicas de *prompting* é a escolha lógica, alinhando-se perfeitamente com nossos objetivos de validar as hipóteses centrais de nossa solução com eficiência e menor custo, dependendo apenas da escolha de um Modelo Fundacional base.

2.2.3 Prompting: como orientar o comportamento do modelo sem treinamento

Na seção anterior, discutimos as diferentes abordagens para implementação de LLMs, com ênfase na escolha do *prompting* devido ao seu potencial de validação de hipóteses sem incorrer em custos e complexidade elevados. Agora, avançamos para examinar como utilizar essa técnica para desenvolver uma aplicação que atenda às necessidades específicas do nosso projeto.

A aprendizagem em contexto (ICT), uma capacidade emergente exclusiva de modelos de grande escala, como introduzido anteriormente, é uma técnica poderosa que pode ser aproveitada através do *prompting*. Essa habilidade foi identificada durante experimentos com modelos a partir de uma grande escala, onde foi observado que, ao fornecer exemplos específicos dentro do próprio prompt, o modelo ajusta sua saída para se alinhar aos padrões desses exemplos, sem a necessidade de alterar seus parâmetros internamente (BROWN et al., 2020). Esta descoberta indicou que modelos complexos podem "aprender" novos comportamentos simplesmente através da exposição a dados de exemplo dentro de uma interação única, o que representa uma vantagem significativa em termos de flexibilidade e aplicabilidade.

Para implementar essa capacidade em nosso MVP, podemos utilizar prompts que incluem exemplos concretos de questões do ENEM, seguidos por respostas-modelo que demonstrem o tipo de comportamento que desejamos que o modelo reproduza. Por exemplo, ao trabalhar com questões de múltipla escolha, o prompt pode apresentar uma série de questões com alternativas, destacando a resposta correta em cada caso, e oferecer ao modelo exemplos do tipo de explicação esperada, por exemplo sem nunca dar a resposta correta "de bandeja" ao usuário. Isso permite que o modelo aprenda a responder de maneira consistente e adequada ao formato esperado, sem a necessidade de ajuste fino.

Outra técnica de *prompting* que será fundamental para o desenvolvimento de nosso MVP é o *Chain-of-Thought Prompting* (CoT). Esta técnica explora a capacidade do modelo de dividir um problema complexo em etapas menores e mais gerenciáveis, facilitando o processo de raciocínio lógico e a geração de respostas coerentes e precisas (BROWN et al., 2020). Para implementar o CoT, podemos estruturar os prompts de maneira que o modelo seja encorajado a "pensar em voz alta" ou detalhar o raciocínio passo a passo antes de chegar à resposta final. No contexto educacional, isso pode ser particularmente útil para ajudar os estudantes a entender não apenas a resposta correta, mas também o processo lógico subjacente à resolução de problemas.

Por exemplo, em uma questão de matemática, o prompt pode instruir o modelo a resolver a questão passo a passo, explicitando cada operação matemática realizada, em vez de apenas fornecer a resposta final. Isso não só melhora a qualidade da resposta, como também fornece aos usuários uma visão detalhada do raciocínio, o que pode ser extremamente benéfico para o aprendizado.

Além disso, a flexibilidade oferecida pelo *prompting* através de APIs permite uma iteração rápida e contínua, onde podemos ajustar e refinar os prompts com base em testes internos, no feedback dos usuários e nos resultados observados. Este ciclo iterativo de ajuste e avaliação é fundamental para o desenvolvimento de um MVP eficaz, pois nos permite explorar diferentes abordagens de *prompting* e identificar aquelas que melhor atendem às necessidades dos estudantes e dos requisitos do projeto.

Em suma, a utilização de técnicas de *prompting*, como a aprendizagem em contexto e o *Chain-of-Thought Prompting*, oferece uma base sólida para o desenvolvimento inicial do nosso MVP. Ao explorar essas técnicas de maneira detalhada e direcionada, podemos criar uma aplicação que não só responde às questões do ENEM de forma eficaz, mas também auxilia os estudantes na compreensão dos conceitos subjacentes, tudo isso sem a necessidade de incorrer em custos elevados ou complexidade adicional.

2.2.4 Recuperação confiável de dados de uma base através de RAG

Retrieval Augmented Generation (RAG) é uma técnica que combina a geração de texto com a recuperação de informações específicas, o que permite que sistemas de IA consultem uma base de dados enquanto geram respostas. No contexto do nosso sistema, o RAG é especialmente relevante para a extração de questões do ENEM de banco com as provas históricas do exame a partir de seu conteúdo, pois as questões são um material fundamental e é essencial garantir que sejam apresentadas em seu formato original. É provável que essas questões tenham sido usadas no pré-treinamento de LLMs como o GPT-4, mas a principal vantagem do RAG é que ele não depende disso, uma vez que esses modelos não são efetivos em recuperar com precisão os textos nos quais foram treinados. Em vez disso, o RAG consulta diretamente uma base de dados específica, como a das provas do ENEM, para extrair informações e gerar respostas baseadas em informações factuais.

O funcionamento do RAG envolve dois componentes principais: o *retriever* e o *generator*. O *retriever* é responsável por buscar, em um conjunto de dados, os documentos mais relevantes

para atender à instrução de um prompt. Geralmente, isso é feito usando técnicas de busca semântica, como o uso de *embeddings* vetorizados das palavras, obtidos através de modelos como BERT (*Bidirectional Encoder Representations from Transformers*) ou Dense Passage Retrieval (DPR). Esses modelos transformam as palavras ou frases em vetores (forma numérica) de alta dimensão que capturam o significado semântico do texto, permitindo que o sistema identifique com precisão os trechos mais relevantes. Uma vez que os documentos relevantes são recuperados, o *generator*, que pode ser um modelo de linguagem como os GPTs, utiliza os trechos recuperados para gerar uma resposta coerente e informativa, incorporando as informações extraídas diretamente do banco de dados (LEWIS et al., 2020).

O acesso a uma base de dados externa é muito útil para nós também por mitigar o problema das alucinações ao ancorar as respostas geradas em dados concretos recuperados com precisão diretamente de uma base de dados. No nosso caso, além de garantir que as questões sejam apresentadas na forma original para os usuários, o próprio modelo ter essa referência ajudará a manter as respostas no escopo definido e atreladas a uma fatualidade validada externamente. O RAG traz a possibilidade de atrelar às questões outros conteúdos que poderão enriquecer as interações com os usuários.

Para maximizar os benefícios do RAG, é fundamental que a base de dados seja organizada e compreensível, o que, no caso do banco de questões do ENEM, já é uma vantagem, dado que as questões são bem estruturadas e categorizadas. Essa organização facilita o processo de recuperação de informações relevantes e minimiza os riscos de selecionar trechos inadequados.

Apesar das vantagens do RAG, sua implementação não está isenta de desafios, especialmente no que diz respeito à sua integração com um sistema recomendador de questões. Embora a base de dados do ENEM seja relativamente simples, a combinação eficiente de RAG com um sistema recomendador exige uma arquitetura específica que possa lidar com a seleção de questões baseadas no histórico do estudante na utilização da solução.

A integração do RAG com sistemas de recomendação envolve selecionar as questões mais apropriadas para o nível de habilidade do estudante, algo que depende fortemente do tipo de algoritmo de recomendação utilizado. Se um sistema recomendador mais simples, baseado em regras, for utilizado, a integração pode ser mais direta, mas menos personalizada. Já sistemas recomendadores mais avançados, como baseados em Teoria de Resposta ao Item, podem oferecer recomendações mais precisas, mas com uma complexidade significativamente maior na integração com o RAG.

Uma abordagem possível para enfrentar esses desafios é a implementação de um pipeline iterativo, onde o sistema recomendador sugere uma lista de questões com base no histórico de

desempenho do estudante, e o RAG refina essas sugestões ao recuperar informações específicas e relevantes para o contexto atual do estudante. Essa abordagem, no entanto, requer uma experimentação intensiva para calibrar os parâmetros e assegurar que as duas partes do sistema funcionem harmoniosamente.

Em última análise, a eficácia da integração do RAG com o sistema recomendador dependerá de testes rigorosos e ajustes contínuos, garantindo que o sistema ofereça uma experiência de aprendizado personalizada e eficaz, minimizando o risco de alucinações e maximizando a relevância das respostas geradas. O uso de um banco de dados não semântico (por exemplo, relacional) para armazenar as questões e um sistema recomendador preciso aliado a prompts efetivos podem dispensar RAG nas versões iniciais da solução, mas esta é uma técnica extremamente relevante para desenvolvimentos futuros de nossa solução.

2.2.5 Um ponto de atenção: as “alucinações” dos LLMs

As "alucinações" dos grandes modelos de linguagem (LLMs) representam um desafio significativo no desenvolvimento de sistemas que dependem da geração de texto com precisão. Alucinações ocorrem quando o modelo gera informações que parecem plausíveis, mas que são, de fato, fictícias ou incorretas. Em nossa aplicação, essa possibilidade é real e preocupante, pois o modelo poderia oferecer respostas falsas aos estudantes, o que não apenas causaria confusão, mas também poderia ir diretamente contra os objetivos de aprimorar a compreensão dos alunos em relação aos conteúdos do ENEM e comprometer o desempenho na prova. Por isso, será necessário testar extensivamente a solução e avaliar especificamente a incidência desse tipo de erro, garantindo que qualquer instância de alucinação seja identificada e mitigada antes do lançamento ao público.

Esse problema não possa ser totalmente prevenido, mas o uso de técnicas como RAG, que permite a recuperação de documentos externos, pode oferecer uma camada adicional de segurança contra alucinações. No nosso caso, pretendemos usa-lo para oferecer ao modelo enunciados, as respectivas alternativas e gabarito (além de outras informações que possam ser desejáveis) com exatidão. Ao fornecer ao modelo o enunciado da questão e o gabarito (ou seja, a alternativa correta), essas informações servem como ponto de partida e de chegada para a geração da resposta, o que em certa medida força o modelo a limitar suas respostas àquilo que é coerente com essas informações. Essa técnica, no entanto, também precisará ser testada e validada para garantir sua eficácia em prevenir alucinações, especialmente em situações onde as informações recuperadas não são suficientemente claras ou detalhadas (JI et al., 2023).

O papel do *prompting* também é crucial nesse contexto, especialmente na tarefa de reforçar a necessidade de se manter preso às informações recuperadas relativas às questões e ao gabarito, bem como ao propósito de oferecer treinamento para a prova. Um desafio adicional é que os LLMs são geralmente treinados para serem solícitos e responder a quaisquer perguntas que não infrinjam padrões de segurança ou impliquem em crimes. Isso significa que, se os usuários fizerem perguntas fora do escopo das questões e da prática para o exame, os modelos podem gerar respostas que extrapolam o conteúdo do ENEM, resultando em informações potencialmente incorretas. Por isso, devemos orientar o modelo a não seguir por esse caminho, limitando-o ao conteúdo específico das questões do ENEM através de técnicas abordadas como ICL e CoT. Simultaneamente, será essencial alertar os usuários de que o uso do modelo para outros fins, que não o treinamento para o ENEM, pode resultar em respostas incorretas e que existe um risco significativo de alucinações.

Do ponto de vista técnico, as alucinações representam um dos principais riscos à viabilidade da nossa solução. No entanto, este é um dos principais focos de estudo de especialistas em LLMs, e continuaremos atentos às novas técnicas e abordagens que possam ser empregadas para mitigar essas falhas. Testes extensivos, acompanhamento constante e ajustes contínuos serão essenciais para garantir que o sistema se mantenha confiável e que suas alucinações sejam minimizadas ao máximo possível.

2.2.6 LLMs são capazes de resolver as questões do ENEM?

Quando foi proposta a criação de um produto de tutoria automatizada para o ENEM, em essência se trata de um software capaz de ajudar estudantes a compreender as questões dos exames passados. Uma etapa anterior e necessária é que tal software seja capaz de compreender as questões do ENEM, afinal, caso contrário, como poderia ajudar? Para isso, verificamos a capacidade de modelos de ponta de resolver as questões e apontar a alternativa correta.

Para manter o rigor, vale notar que compreender uma questão e apontar a alternativa correta não são exatamente o mesmo, já que é possível acertar uma resposta sem nem ler a questão. Tomamos por premissa, porém, que não existe qualquer outro padrão além da compreensão das relações lógicas e de sentido entre o enunciado e as alternativas que possa ser aprendido e que conduza até a resposta correta de cada questão. De maneira que a única forma de obter consistentemente respostas corretas, ou seja, obtendo uma taxa de acerto elevada em

uma longa série de respostas, é por meio da compreensão. Considerando um modelo de questionário de múltipla escolha com 5 respostas alternativas para cada enunciado, há uma chance de 20% de optar pela resposta correta ao escolher uma por acaso, e a probabilidade para uma sequência de $n + m$ questões com n acertos e m erros é dada por:

$$\text{comb}(n + m | n) \times (0.2)^n \times (0.8)^m$$

Admitindo que as únicas opções de acerto são a compreensão ou o acaso, para uma dada extensão de uma sequência de questões e uma dada taxa de acerto, podemos verificar a probabilidade das respostas terem sido dadas ao acaso e determinar se consideramos essa possibilidade factível diante de um *threshold* escolhido.

Existem alguns trabalhos que verificaram a capacidade de LLMs de ponta em resolver o ENEM. Nunes et al. (2023) testaram o desempenho do GPT-3.5 e GPT-4 nos ENEMs entre 2009-2017 e de 2022 e obtiveram resultados muito positivos. Pires et al. (2023) pesquisadores da empresa Maritaca.ia, que produz LLMs nativos para o português brasileiro, exploraram as capacidades de visão do GPT-4 para responder questões dependentes de compreensão de imagens nos ENEMs de 2022 e 2023. Almeida et al. (2024) também da Maritaca.ai, exploraram no relatório técnico da sua família de modelos Sabiá-2 o desempenho no ENEM e outros exames padronizados em português brasileiro, constatando a versão Sabiá-2 Medium ter uma performance superior ao GPT-3.5 e inferior ao GPT-4 na maioria dos exames, destacando as limitações do modelo para questões de matemática e programação.

Nunes et al. (2023) testaram o GPT-3.5 e GPT-4 primeiramente no *dataset* “ENEM Challenge”, que contém questões do exame entre os anos de 2009 e 2017, detalhado na seção 4. Para avaliar os modelos selecionados, os autores removeram questões classificadas como sendo de Compreensão de Imagem, Raciocínio Matemático e Elementos Químicos (que contém figuras de compostos químicos). Os resultados neste *dataset* e no de 2022 com *zero-shot prompting* estão na tabela 1.

Tabela 1 – Resultados no ENEM 2009-2017 com *zero-shot prompting*.

Model	2009	2010	2011	2012	2013	2014	2015	2016_1	2016_2	2017	avg.	2022
Silveira et. al	26.96	31.37	33.85	31.34	30.33	27.01	29.96	28.42	27.95	26.68	29.39	-
text-davinci-002	74.16	77.45	80.21	78.26	82.35	82.76	80.90	77.66	74.19	79.78	78.77	84.09
text-davinci-003	68.54	72.55	83.33	78.26	77.65	79.31	76.40	79.79	78.49	79.78	77.41	79.55
code-davinci-002	83.15	82.35	87.50	81.52	80.00	87.36	84.27	75.53	81.72	85.39	82.88	85.23
gpt-3.5-turbo	78.65	77.45	85.42	84.78	74.12	82.76	84.27	85.11	76.34	84.27	81.32	86.36
gpt-4	95.51	92.16	97.92	98.91	91.76	96.55	96.63	92.55	89.25	94.38	94.56	90.91

Fonte: Nunes et al (2023).

A avaliação dos autores teve como resultado um desempenho bom para os modelos da família GPT-3.5 (code-davinci-002, text-davinci-002, text-davinci-003, gpt-3.5-turbo) e GPT-4, com uma taxa de acerto de respectivamente 81.32% e 94.56% entre os anos de 2009-2017, e de 86.36% e 90.91% no ENEM 2022. Pode-se notar que a pior performance do GPT-4 foi no ENEM 2022, e em relação a isso os autores não descartam que tenha havido contaminação dos dados: como os modelos haviam no momento do estudo sido treinados com dados até setembro de 2021, é possível (e até provável) terem sido expostos às questões do ENEM de 2009 até 2017 durante a etapa de pré-treino. Ainda que dificilmente arquivos que relacionam as questões diretamente às respostas tenham sido utilizados, entende-se que os resultados para o ENEM de 2022 são mais confiáveis, e 86.36% e 90.91% são bastante altos.

Para as 118 questões deste dataset, uma taxa de acerto de 90.91% implicaria numa probabilidade $\text{comb}(118 \times (118 \times 0.9091)) \times (0.2)^{(118 \times 0.9091)} \times (0.8)^{(118 \times 0.0909)} = 1.33 \times 10^{-61}$, um número bastante irreal independente do *threshold* escolhido, o que leva a concluir que os LLMs GPT-3.5 e GPT-4 tem capacidades significativas de compreensão das questões do ENEM.

Tabela 2 – Resultados no ENEM 2022.

Area	code-davinci-002			gpt-3.5-turbo			gpt-4		
	zero-shot	three-shot	three-shot with CoT	zero-shot	three-shot	three-shot with CoT	zero-shot	three-shot	three-shot with CoT
Languages and Codes	78.79	87.88	72.73	75.76	81.82	69.70	84.85	87.88	87.88
Human Sciences	89.19	94.59	91.89	91.89	89.19	94.59	94.59	94.59	94.59
Natural Sciences	69.23	61.54	53.85	73.08	84.62	65.38	84.62	76.92	88.46
Mathematics	18.18	27.27	50.00	18.18	36.36	54.55	40.91	50.00	72.73
Total	68.64	72.88	70.34	69.49	76.27	73.73	79.66	80.51	87.29

Fonte: Nunes et al (2023).

Na tabela 2, Nunes et al. (2023) apresentam em maiores detalhes os resultados obtidos no ENEM 2022 em que apenas questões de compreensão de imagem foram removidas. É relevante para nós observar a diferença no desempenho dos modelos entre diferentes áreas do conhecimento com *zero-shot prompting*. Percebe-se uma queda significativa de performance em matemática em relação às outras 3 categorias (línguagens, ciências humanas e ciências naturais). Para os modelos da família GPT-3.5, a taxa de acerto em matemática foi de apenas 18.18%, abaixo dos 20% esperados caso o modelo não tivesse compreensão alguma, e para o GPT-4, 40.91%, que ainda não é nada próximo de um resultado satisfatório caso fossemos avaliar usar esse modelo e *zero-shot prompting* como fundamento para a tutoria de questões de matemática do ENEM em nosso produto. Em relação às outras categorias, temos para o GPT-

3.5 turbo e GPT-4 um desempenho similar entre linguagens e ciências naturais, respectivamente por volta de 74% e 84% de acertos, e um desempenho excelente para ciências humanas, respectivamente 89.19% e 94.59%.

O outro ponto muito esclarecedor do estudo em questão é a diferença de performance que ele demonstra de acordo com diferentes técnicas de *prompting*. Sobretudo nas questões de matemática, técnicas de *few-shot* permitiram melhoria de 200% no GPT-3.5 e 77% no GPT-4, alcançando inclusive um nível de acerto razoável no caso do último. No geral, o GPT-4 melhorou quase 10% a partir de instruções dadas a ele, o que nos mostra que LLMs não tem uma capacidade estática de compreensão das questões, mas que a forma como são instruídos a abordar problemas tem um efeito totalmente significativo.

Em relação ao uso de CoT, os autores nos mostram que instruir o modelo a decompor as questões em etapas sequenciais leva a um resultado melhorado, sobretudo em matemática. Ao realizar essa decomposição, o modelo demonstra inclusive uma capacidade para explicar as etapas intermediárias de resolução através de linguagem, algo fundamental tendo em vista a funcionalidade de explicar questões de matemática para os usuários da nossa solução.

Na imagem abaixo, vemos exemplos de questões de física e matemática, respectivamente, em que o GPT-4 errou sem usar o recurso do CoT mas acertou ao usar. Em ambos os casos, foi capaz de dividir a solução em etapas lógicas, recuperar os conceitos necessários em cada etapa (no caso fórmulas) e aplicá-los corretamente, chegar a resposta correta e explicar de maneira comprehensível cada processo. Para isso, partiu de exemplos de resoluções de outras questões com etapas e conceitos diversos, mostrando que a técnica de CoT permitiu desbloquear capacidades que não estava empregando sem ela.

Figura 1 – Comparaçao de respostas geradas por GPT-4 sem e com CoT

<p>Question</p> <p>In 2017, the Parque Solar Lapa was inaugurated in the state of Bahia, consisting of two plants (Bon Jesus da Lapa and Lapa) and capable of generating around 300 GWh of energy per year. Assume that each plant has a capacity equal to 75 MW, with the park totaling an installed capacity of 150 MW. Also consider that the average solar irradiance is 1 500 W/m² and that the efficiency of the panels is 20%. Under these conditions, the total area of the solar panels that make up Parque Solar Lapa is closer to:</p> <p>A. 1 000 000 m² B. 500 000 m² C. 250 000 m² D. 100 000 m² E. 20 000 m²</p> <p>3-shot</p> <p>D. X</p> <p>To calculate the total area of solar panels, we first need to find the total power generated by the panels. We know that the installed power of the park is 150 MW, that is, 150,000,000 W. As the efficiency of the panels is 20%, the power generated by the panels is 150,000,000 W * 0.20 = 750,000,000 W.</p> <p>Now, we need to find the area needed to generate this power, considering an average solar irradiance of 1,500 W/m². To do this, just divide the power generated by the average solar irradiance:</p> <p>Area = Power generated / Average solar irradiance Area = 750,000,000 W / 1,500 W/m² Area = 500,000 m²</p> <p>Therefore, the total area of the solar panels that make up Parque Solar Lapa is closer to 500,000 m².</p> <p>Response: B.</p>	<p>Question</p> <p>A cook makes special sweets for delivery. Using a base dough recipe, she prepares a portion, with which she produces 50 solid spherical sweets, 2 cm in diameter. A customer orders 150 of these sweets, but asks that each one be spherical in shape, 4 cm in diameter. The cook intends to prepare the exact number of servings of the dough base recipe needed to produce the sweets for that order.</p> <p>How many servings of the base pasta recipe should she prepare to serve this customer?</p> <p>A. 2 B. 3 C. 6 D. 12 E. 24</p> <p>3-shot</p> <p>C. X</p> <p>First, let's calculate the volume of a candy with a diameter of 2 cm. The radius is half the diameter, so $r = 1$ cm. The volume of a sphere is given by $V = \frac{4}{3}\pi r^3$. So, the volume of a 2 cm diameter candy is $V = \frac{4}{3}\pi(1)^3 = \frac{4}{3}\pi\text{cm}^3$.</p> <p>Now, let's calculate the volume of a candy with a diameter of 4 cm. The radius is half the diameter, so $r = 2$ cm. The volume of a sphere is given by $V = \frac{4}{3}\pi r^3$. So, the volume of a 4 cm diameter candy is $V = \frac{4}{3}\pi(2)^3 = (32/3)\pi\text{cm}^3$.</p> <p>The cook manages to make 50 sweets of 2 cm in diameter with one portion of the basic recipe. So the total volume of dough she can make from one serving is $50 * (4/3)\pi\text{cm}^3$. The customer ordered 150 sweets of 4 cm in diameter. So, the total volume of dough required for this order is $150 * (32/3)\pi\text{cm}^3$.</p> <p>To find out how many servings of the base recipe you need, simply divide the total volume of dough needed by the total volume of dough she can make in one serving: $(150 * (32/3)\pi\text{cm}^3) / (50 * (4/3)\pi\text{cm}^3) = (150 * 32) / (50 * 4) = 4800 / 200 = 24$.</p> <p>Therefore, the cook must prepare 24 servings of the basic pasta recipe to serve this customer.</p> <p>Response: E.</p>
---	--

(a) Question 134 of Natural Sciences

(b) Question 141 of Mathematics

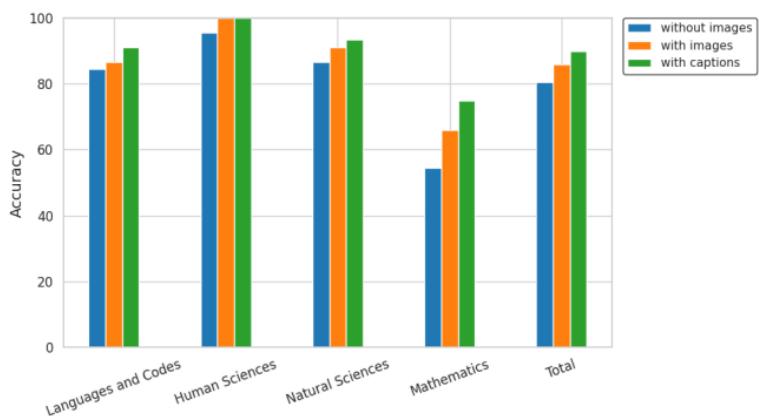
Fonte: Nunes et al (2023).

Exemplos e resultados como estes nos deixam otimistas com as perspectivas de instruir o modelo a fornecer explicações valiosas para os estudantes, e sugerem o GPT-4 como uma boa opção de Modelo Fundacional a ser utilizado no MVP.

Em um estudo similar realizado por pesquisadores da Maritaca.ai, Pires et al., 2023, investigaram a efetividade das capacidades de visão de modelos da família GPT-4 na resolução de questões do ENEM. Aqui não foram removidas questões, e os autores avaliaram os modelos em relação a todas as questões dos ENEMs de 2022 e 2023 (este sem risco de contaminação de dados).

Foram testadas 3 abordagens diferentes: sem as imagens, com as imagens e com as descrições textuais usadas nos exames com recursos de acessibilidade, e para todas as matérias o uso de imagens se provou superior ao não uso, e ambas inferiores ao uso de descrições textuais. Infelizmente os autores não apresentaram as diferenças de resultado para apenas as questões que envolviam compreensão de imagem, não sendo possível derivar do estudo uma taxa de melhora neste tipo específico de questão.

Figura 2 – Performance do GPT-4 no ENEM 2023



Fonte: Pires et al (2023).

Estes resultados são de nosso interesse pois o ENEM é um exame que utiliza-se bastante de recursos visuais, e lidar com imagens é uma dificuldade antevista, já que depende de um outro tipo de *dataset* e da utilização de modelos com capacidade de visão. Por conta dessas dificuldades, não utilizaremos as questões com imagens no nosso MVP. Mas os resultados apresentados por Pires et al. são animadores com relação às perspectivas futuras de incorporar essas questões, ainda mais levando em conta a curva de aperfeiçoamento dos LLMs de ponta. Modelos multimodais como o GPT-4o demonstram uma capacidade muito maior de lidar com imagens, e também com raciocínio matemático.

Em relação a todos os resultados observados, vale apontar que não contém variabilidade estatística e que nos são úteis sem depender de um grau de precisão alto para validar a compreensão por LLMs de ponta com relação às questões do ENEM.

2.3 Teoria de Resposta ao Item (TRI)

2.3.1 Adequação à recomendação de questões visando uma prática otimizada para estudantes individuais

A Teoria de Resposta ao Item (TRI), desenvolvida inicialmente na primeira metade do século XX, é uma poderosa ferramenta para a análise de dados em testes padronizados e é hoje uma metodologia amplamente aceita na mensuração de habilidades latentes em contextos educacionais. Diferente da Teoria Clássica dos Testes, que se baseia em pressupostos como a linearidade entre a habilidade do indivíduo e a pontuação obtida, a TRI oferece um modelo mais robusto e matematicamente sofisticado, permitindo uma análise mais precisa e flexível das respostas dos indivíduos (LORD, 1980). A grande inovação da TRI está na sua capacidade de modelar a probabilidade de uma resposta correta a um item específico como uma função tanto das características do item quanto das habilidades do respondente, proporcionando uma compreensão mais profunda e detalhada dos processos subjacentes à realização de testes educacionais.

No âmago da Teoria de Resposta ao Item está a modelagem da relação entre a habilidade do indivíduo, frequentemente denominada como traço latente, e a probabilidade de uma resposta correta a um item. Essa relação é descrita pela função de resposta ao item, que se baseia em três parâmetros principais: dificuldade, discriminação e adivinhação. O parâmetro de dificuldade indica o nível de habilidade necessário para ter 50% de probabilidade de responder corretamente a um item; já o parâmetro de discriminação mede o quão bem o item distingue entre indivíduos com habilidades próximas. Por fim, o parâmetro de adivinhação representa a probabilidade de acertar o item por pura sorte, especialmente relevante em questões de múltipla escolha. Esses parâmetros juntos formam a curva característica do item (CCI), uma representação gráfica que captura a complexa interação entre a habilidade do respondente e a probabilidade de acerto, oferecendo uma visão mais rica e detalhada da qualidade de cada item do teste (EMBRETSON; REISE, 2000).

A aplicação da IRT em testes padronizados, como o ENEM, permite uma avaliação mais precisa e equitativa das habilidades dos estudantes. Ao considerar as características individuais de cada item e as habilidades dos respondentes, a TRI facilita a construção de testes que são capazes de medir habilidades latentes de forma independente da amostra de indivíduos ou do conjunto específico de itens utilizados (VAN DER LINDEN; HAMBLETON, 1997). Isso é particularmente importante em um país como o Brasil, onde as desigualdades socioeconômicas podem introduzir vieses significativos na avaliação educacional. A TRI, ao permitir a comparação direta entre diferentes versões do exame, é uma tentativa de medir a proficiência de maneira mais consistente e justa, independentemente das circunstâncias específicas de cada aplicação do teste. Além disso, a metodologia da TRI apoia a implementação de sistemas adaptativos, onde o nível de dificuldade dos itens apresentados aos estudantes pode ser ajustado em tempo real, com base em suas respostas anteriores, tornando o processo de avaliação não apenas mais preciso, mas também mais personalizado e eficiente; este é o uso que nos interessa mais aqui.

Uma das principais aplicações da TRI em produtos educacionais ocorre a partir da sua capacidade de personalizar o ensino, adaptando a prática dos estudantes ao seu nível de habilidade individual. Ao utilizar a TRI para analisar as respostas dos estudantes, é possível identificar não apenas quais áreas necessitam de maior atenção, mas também quais itens ou tipos de questões são mais adequados para promover o aprendizado. Esse tipo de análise é muito útil para a construção de sistemas de tutoria automatizados, que podem recomendar questões que maximizem o potencial de aprendizado dos estudantes, oferecendo uma experiência de estudo mais direcionada e eficiente (WEISS, 1983). No contexto do ENEM, a utilização da TRI pode transformar a preparação para o exame, permitindo que os estudantes pratiquem com questões que são precisamente ajustadas ao seu nível atual de conhecimento, o que poderá ser um elemento muito importante para aumentar a probabilidade de sucesso nas provas.

2.3.2 Ponderações sobre a implementação de um sistema recomendador de questões do ENEM a partir da TRI

A implementação da Teoria de Resposta ao Item em sistemas de ensino adaptativos tem como condição básica a manutenção precisa dos parâmetros dos itens - dificuldade, discriminação e adivinhação. Esses parâmetros formam a base para as decisões de recomendação de conteúdo, e sua precisão é fundamental para garantir que o sistema adaptativo possa fornecer recomendações eficazes e personalizadas aos estudantes.

No contexto do ENEM, uma abordagem viável seria a solicitação formal dos parâmetros de TRI ao Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), que poderia fornecer dados já validados. Esses dados são essenciais para assegurar que o sistema se beneficie de uma base robusta e confiável, permitindo que as recomendações sejam feitas com precisão desde as primeiras versões da solução.

Entretanto, caso os parâmetros do INEP não sejam obtidos, seria necessário estimar esses parâmetros do 0 com base nas respostas dos próprios usuários do sistema. Isso apresenta desafios significativos, especialmente considerando a necessidade de um número substancial de respostas por questão para calcular os parâmetros de forma precisa. Estudos sugerem que são necessárias aproximadamente 200 a 500 respostas por questão para estimar de forma confiável os parâmetros da TRI (LORD, 1980). Considerando que a base histórica do ENEM pode conter cerca de 180 questões por ano, e assumindo a inclusão de várias edições do exame para garantir variabilidade, seria necessário coletar dezenas de milhares de respostas antes de atingir um nível mínimo de precisão para o sistema recomendador.

Dado o volume de respostas necessário, fica claro que a coleta desses dados a partir de interações iniciais com usuários pode levar a uma imprecisão significativa nos parâmetros da TRI, especialmente nas fases iniciais do MVP. Por isso, a obtenção dos parâmetros diretamente do INEP seria uma condição importante para o uso da TRI nas primeiras versões da solução. Sem esses dados, a implementação de um sistema recomendador baseado em TRI seria altamente limitada.

A implementação prática de sistemas adaptativos baseados em TRI também exige uma infraestrutura de dados minimamente sofisticada, capaz de processar dados em tempo real, como as respostas dos estudantes, para ajustar as recomendações conforme o desempenho individual. Esse sistema deve suportar a complexidade de monitorar continuamente o progresso dos estudantes e adaptar o conteúdo de forma dinâmica, garantindo uma experiência de aprendizado personalizada e eficaz. A complexidade da infraestrutura necessária dependerá do número de usuários e da frequência de uso.

Do ponto de vista matemático, a aplicação da TRI em sistemas adaptativos envolve a utilização de modelos estatísticos avançados para estimar os parâmetros dos itens e as habilidades dos estudantes. Esses modelos são intrinsecamente complexos e demandam uma abordagem cuidadosa, que leve em conta as variações nas respostas dos estudantes e ajuste os parâmetros de forma contínua. A precisão na calibração desses modelos é essencial para assegurar que o sistema adaptativo funcione de maneira eficaz e ofereça recomendações que realmente atendam às necessidades individuais dos estudantes.

Embora a implementação detalhada desse modelo matemático esteja fora do escopo do presente trabalho, é importante reconhecer que o domínio e a aplicação eficaz dessa etapa são componentes essenciais para elevar a solução proposta a um novo nível de sofisticação e eficácia. A complexidade técnica e computacional envolvida na calibração contínua e na personalização em tempo real, bem como na avaliação de sua efetividade, reforça a necessidade de futuras colaborações interdisciplinares, que poderiam complementar e expandir as capacidades do sistema adaptativo proposto.

2.3.3 Alternativas de sistemas recomendadores

Diante das limitações para implementação de um sistema recomendador baseado na Teoria de Resposta ao Item sem o suporte de dados externos ou um volume significativo de respostas dos usuários, é importante explorar alternativas que possam oferecer recomendações com eficácia desde as primeiras fases do projeto.

A consideração do nível de dificuldade das questões, fundamental para um sistema recomendador robusto, enfrenta desafios substanciais nessas condições. Mas esses obstáculos podem ser contornados através de soluções criativas que permitem uma personalização inicial, mesmo em um cenário de dados limitados.

Uma dessas possibilidades é a personalização via formulários de entrada, onde os usuários fornecem informações sobre suas habilidades, áreas de interesse e tópicos que consideram mais desafiadores. Ao coletar esses dados, o sistema pode construir um perfil inicial de cada usuário, permitindo a personalização das recomendações de forma imediata, sem depender de interações prévias no sistema.

Essa abordagem pode ser complementada de forma inovadora com o uso de LLMs para analisar o conteúdo das questões, avaliar os tópicos envolvidos nas questões e estimar seus níveis de dificuldade. Essa abordagem pode ser usada inclusive para fornecer uma primeira aproximação a parâmetros psicométricos, como a dificuldade. Embora esses parâmetros iniciais sejam menos precisos do que aqueles derivados de grandes volumes de dados de usuários, à medida que mais respostas são coletadas, essas aproximações iniciais poderão ser ajustadas para refletir melhor a realidade, com uma melhora gradual da precisão das recomendações.

3 MODELO DE EXECUÇÃO E REFERENCIAL PRÁTICO

Tendo coberto aspectos teóricos que direcionam e fundamentam a solução, nesta seção o foco é o estudo do ambiente de negócios, envolvendo fatores que condicionam sua realização efetiva. Primeiro, serão feitas considerações sobre o modelo de negócios, com aprofundamento da perspectiva B2G, e uma breve abordagem sobre as possibilidades B2B e B2C. Em seguida, trataremos do mercado brasileiro de soluções tecnológicas voltadas à preparação para o vestibular, numa busca de situar a proposta em relação ao que já existe e agregar o máximo de ensinamentos que for possível. Então, alguns elementos serão trazidos para uma discussão sobre o momento dos LLMs em relação ao mercado e soluções educacionais baseadas neles, com especial atenção para aquelas voltadas a tutoria. Finalmente, será apresentada uma breve análise estratégica, com aspectos chave trazidos por metodologias como a matriz SWOT e o *business Canvas*.

3.1 Considerações sobre o modelo de negócios

Considerando o contexto educacional brasileiro descrito na introdução deste trabalho e os fundamentos teóricos que embasam essa proposta, acredita-se que um sistema de tutoria automatizado baseado em LLMs voltado à preparação para o ENEM é não apenas viável do ponto de vista técnico, mas também capaz de gerar um impacto significativo tanto para os indivíduos quanto para a sociedade. Segundo as teorias e experimentos discutidos, a tutoria é um método de ensino que maximiza o aprendizado ao proporcionar um ambiente de estímulo constante e ao adaptar o fluxo de atividades e seus conteúdos às características particulares de cada estudante. Com o avanço das tecnologias de IA, especialmente os LLMs, surge pela primeira vez a possibilidade de criar sistemas de tutoria escaláveis, acessíveis e capazes de interagir com os alunos de maneira personalizada, utilizando a linguagem natural. Essa inovação abre caminhos para a democratização da educação de qualidade, oferecendo suporte educacional eficaz a um custo muito menor do que os métodos tradicionais.

No entanto, viabilizar essa solução não é uma tarefa simples. Embora os modelos fundacionais, como os LLMs, ofereçam grande flexibilidade, o desafio reside em transformá-los em um sistema engajador, que os estudantes possam utilizar de forma autônoma. Além disso, é crucial programar a solução para que ela se adapte ao progresso individual de cada usuário, o que envolve recomendar conteúdos de maneira precisa e garantir a consistência em interações que sejam verdadeiramente instrutivas. O objetivo é encontrar um equilíbrio entre a

flexibilidade inerente dos LLMs e a consistência necessária para fins educacionais, afinando o sistema para maximizar seu potencial instrutivo. Uma maneira de alcançar esse equilíbrio é por meio de intensa experimentação e interação com usuários reais, um processo que será abordado na Seção 4, com a descrição de experimentos e especificação do MVP.

Para viabilizar esse processo, é essencial desenvolver um modelo de negócios sustentável. Os recursos necessários para iniciar o processo com as primeiras validações serão relativamente modestos, já que existem soluções pré-formatadas que nos permitem avançar rapidamente para testar aspectos chave da proposta. No entanto, para desenvolver uma solução capaz de entregar valor real aos usuários estudantes e, sobretudo sustentar os processos contínuos de experimentação e refinamento, será necessário um investimento significativo ao longo de um período prolongado. Neste contexto, serão exploradas diferentes alternativas de modelos de negócios, sempre com foco nos objetivos de geração de valor a serem alcançados.

Dado o objetivo de gerar impacto social no âmbito da educação pública, o modelo B2G (*Business-to-Government*) se apresenta como o mais adequado. Dentro da proposta de oferecer recursos educacionais que possibilitem, especialmente aos estudantes em condições socioeconômicas desfavoráveis, competir em igualdade de condições por vagas nas universidades, ao lado de seus pares mais privilegiados, a principal dificuldade reside em alcançar uma parcela significativa desses estudantes. Somente no ENEM 2023, 3,9 milhões de candidatos participaram da prova (INEP, 2023). O governo, além de ser o principal interessado em uma solução que possa reduzir desigualdades, conforme previsto na Constituição Federal, possui acesso aos canais de distribuição mais eficazes: espaço na mídia impressa e audiovisual, e, sobretudo, nas escolas públicas de ensino básico, onde se concentra nosso público-alvo. A prática de investimento público em projetos inovadores ainda não é muito difundida no Brasil, mas isso não significa que inexistem oportunidades de financiamento público no país, e estes podem ser cruciais para a sustentação do projeto. Há algumas fontes de financiamento que podem ser acessadas para projetos que alinhjam tecnologia e educação com impacto social.

O Fundo Nacional de Desenvolvimento da Educação (FNDE) é uma das principais fontes de recursos para iniciativas educacionais no país. Este fundo, que administra um orçamento de mais de R\$ 5,4 bilhões, é voltado majoritariamente para a educação pública, com programas que promovem a inclusão digital e a inovação tecnológica. Embora o FNDE seja tradicionalmente direcionado a instituições públicas, parcerias estratégicas com essas entidades podem permitir que iniciativas privadas, como a nossa, também acessem esses recursos. Programas como o ProInfo, que financia a implementação de tecnologia nas escolas públicas, poderiam ser um caminho viável para colaborar com o FNDE e obter financiamento.

Outro fundo essencial é o Fundo Nacional de Desenvolvimento Científico e Tecnológico (FNDCT), gerido pelo Ministério da Ciência, Tecnologia e Inovações (MCTI). Este fundo, que em 2023 contou com um orçamento de R\$ 10 bilhões, é destinado a apoiar projetos que promovam a inovação científica e tecnológica com relevância social. O FNDCT frequentemente abre editais para a participação de entidades privadas, especialmente aquelas que desenvolvem projetos com impacto significativo no campo da ciência e tecnologia. A proposta de um sistema de tutoria automatizado, que utiliza tecnologias de ponta para democratizar o acesso à educação, se alinha com os objetivos do FNDCT, o que faz deste fundo uma excelente oportunidade de financiamento.

O Banco Nacional de Desenvolvimento Econômico e Social (BNDES) também se destaca como uma fonte importante de recursos. O BNDES oferece diversas linhas de crédito e financiamento para projetos de inovação, incluindo aqueles voltados para a educação. Recentemente, o banco tem intensificado seu apoio a iniciativas que combinam impacto social e inovação tecnológica. Com programas que priorizam projetos em regiões com baixo Índice de Desenvolvimento Humano (IDH) e que promovem a inclusão digital e social, o BNDES oferece um caminho robusto para captar recursos necessários à implementação e escalabilidade do nosso projeto.

Por fim, o programa Universidade Aberta do Brasil (UAB), focado na educação a distância, oferece suporte financeiro a projetos que visam ampliar o acesso ao ensino superior. Embora o financiamento da UAB seja direcionado principalmente a instituições públicas, colaborações com universidades federais podem abrir oportunidades para integrar nossa solução ao sistema público de ensino superior, potencializando seu alcance e impacto.

Para maximizar as chances de sucesso na captação desses recursos, é imprescindível acompanhar de forma contínua os editais e chamadas públicas dos órgãos financiadores, submetendo propostas que não apenas evidenciem a relevância social e a viabilidade técnica do projeto, mas que também estejam claramente alinhadas com as diretrizes das políticas públicas educacionais do Brasil. Em um cenário competitivo, a diferenciação de nossa proposta depende de um planejamento detalhado e de uma comunicação clara do potencial transformador que o projeto pode trazer ao sistema educacional brasileiro. No entanto, ao lidar com um projeto altamente inovador, que naturalmente carrega incertezas, torna-se ainda mais crucial adotar uma abordagem integrada que vá além de um plano de negócios detalhado. Isso inclui a incorporação de ensinamentos extraídos de iniciativas similares, que possam oferecer insights valiosos sobre práticas eficazes e potenciais armadilhas, bem como a evolução contínua do projeto para validar as hipóteses mais incertas das quais seu sucesso depende. Ao agregar essas

práticas e desenvolver uma estratégia de mitigação de riscos, é possível tornar essa proposta mais robusta aos olhos dos tomadores de decisão e aumentar as chances de obtenção do suporte necessário para que os impactos pretendidos sejam efetivamente alcançados.

Apesar do contexto de alta imprevisibilidade no que diz respeito ao acesso a recursos públicos, é possível avançar com o projeto sem esse tipo de financiamento. Uma estratégia viável é iniciar com capital privado e desenvolver um MVP que possa ser testado com usuários reais, que aumentará as chances de financiamentos (públicos ou não) posteriores. Para isso, podem-se explorar as possibilidades oferecidas pelos modelos B2C (*Business-to-Consumer*) e B2B (*Business-to-Business*).

Embora não haja perspectivas animadoras de viabilizar a proposta visando diretamente os consumidores finais, devido ao desafio de alcançar uma escala suficientemente grande para tornar a solução de baixo custo acessível a um público com limitações financeiras, essa estratégia ainda pode ser valiosa nas etapas iniciais. A criação de um MVP funcional, com recursos limitados, permitirá validar a proposta de valor ao testar a disposição dos usuários em pagar, mesmo que um valor simbólico, pelo serviço. Além disso, o lançamento desse MVP para um público inicial pode ser realizado por meio de diversas estratégias, como parcerias com escolas, engajamento de influenciadores em redes sociais ou mesmo campanhas modestas de mídia paga nas redes, segmentadas especificamente para o público-alvo desejado.

Uma estratégia similar, mas voltada para a venda direta às escolas, pode ser atraente do ponto de vista do financiamento inicial. Ao atuar como intermediárias entre nossa solução e um número maior de alunos, as escolas podem ampliar significativamente as possibilidades de receita, embora essa abordagem exija um processo de vendas mais complexo. Nessa etapa, a validação da proposta pode ser realizada com escolas privadas como clientes iniciais, que podem perceber nossa solução como uma vantagem competitiva para melhorar os resultados dos alunos nos vestibulares, a um custo relativamente baixo. Levar adiante essa estratégia não só possibilita a evolução do produto até um estágio de maior maturidade, como também pode atrair o interesse de investidores privados, abrindo portas para futuras expansões.

Uma outra alternativa que pode ser extremamente interessante é buscar recursos por meio de editais privados promovidos por fundações ou instituições dedicadas à filantropia. Organizações como a BrazilFoundation, Instituto Ayrton Senna, Fundação Itaú, Fundação Lemann e Instituto Embraer são exemplos de entidades que frequentemente oferecem financiamento para projetos com forte impacto social e inovador. No Anexo A é apresentado o edital da Fundação Itaú “IA para Educação”, que serve de exemplo. Além de fornecer os recursos necessários para desenvolver e amadurecer nossa solução, essas instituições mantêm

interfaces significativas com o poder público, o que pode facilitar a implementação e a escalabilidade da proposta.

Com essas considerações em mente, é essencial voltar a atenção para o mercado brasileiro de soluções baseadas em tecnologia voltadas à preparação para o vestibular, com foco especial no ENEM. A próxima seção abordará uma análise das opções já disponíveis no mercado, examinando suas características, similaridades e diferenças. Essa pesquisa é fundamental para situar nossa solução dentro do panorama atual, permitindo que possamos identificar tanto as lacunas a serem preenchidas quanto as práticas que já se mostraram bem-sucedidas. Ao integrar esses elementos comprovados ao nosso projeto, aumentamos significativamente nossas chances de sucesso e fortalecemos a confiança de possíveis financiadores, tanto no setor público quanto no privado, ao demonstrar que nossa proposta está enraizada em realidades testadas e validadas no contexto educacional brasileiro.

3.2 Panorama do mercado brasileiro de soluções para preparação ao vestibular baseadas em tecnologia

Historicamente, a demanda por preparação para vestibulares foi suprida por instituições que oferecem cursos presenciais e materiais de confecção própria, os cursinhos com suas apostilas proprietárias. Primeiro eram especializadas em exames específicos e muito concorridos, exemplos são as faculdades de medicina, ou então grandes vestibulares como a FUVEST e concursos para vagas federais bem remuneradas. Não foram obtidos dados sobre estes setores, mas diversos documentos apontam o aumento de relevância dos cursinhos ao longo do tempo. Em relação à preparação para vestibulares, esse crescimento ganhou força a partir da introdução do ENEM e aumento no número de vagas em universidades federais. Em 2010, por volta de 48.000 vagas em universidades federais foram distribuídas pelo SISU, em oposição a 240.000 vagas em 2018, um crescimento de 400% em 8 anos (MEC, 2019).

Esse mercado vem se transformando profundamente a partir das EdTechs, startups do setor educacional. Startups são empresas de tecnologia que implementam soluções de alta escalabilidade, sendo uma categoria de organizações que navegam em cenários de muita incerteza e que se veem obrigadas a mitigar riscos para aumentar suas chances de sucesso. Segundo relatório da ABStartups de 2022, neste ano o Brasil contava com 813 EdTechs, sendo que 82% haviam sido fundadas após 2016. Em média, estas empresas haviam recebido R\$1.316.116,50 de investimento desde investimento anjo até série A e acima. 98% ofereciam

solução por meio de software e 12% pertenciam à categoria de cursos preparatórios. Dessas, 17 startups foram analisadas: 47% já haviam recebido investimentos e, destas, 50% após 2020 (apenas entre 2020 e 2022). Em 2022, o mercado global de EdTechs foi avaliado em US\$123,40 bilhões e, para 2023, a expectativa é de que alcance US\$142,37 bilhões. A previsão é que o setor cresça a uma taxa anual de 13,6% entre 2023 e 2030 (TOTVS, 2022). Esses números revelam o panorama de um mercado extremamente aquecido para aplicações de tecnologia a educação, com muitas empresas sendo fundadas e disponibilidade de capital, além da previsão de crescimento.

A plataforma de inteligência de mercado da economia de impacto HolonIQ realiza anualmente desde 2020 o relatório anual “The Latin America EdTech 100”, em que aponta as 100 startups mais promissoras da América Latina (o relatório desconsidera empresas fundadas há mais de 10 anos). O Brasil aparece como amplamente dominante, com 39 empresas apontadas em 2021 e 2022, e 42 em 2023; neste ano seguido pelo México com 18 e pela Argentina com 13. Dentre os subsegmentos apontados no relatório, aquele que nos interessa mais é de “*Tutoring and Test Preparation*”, focado em ensino com características individualizadas e de preparação para exames. Todas as empresas brasileiras dessa categoria que foram apontadas nas 4 edições do relatório e que tem foco na preparação para vestibulares ou em tutoria automatizada são apresentadas na tabela 3.

Tabela 3 – Comparativo de soluções tecnológicas de preparação ao vestibular

	Descomplica	TutorMundi	PasseiDireto
Produto	Plataforma EAD	Tutoragem sob demanda	Plataforma de conteúdo
Modelo	B2C	B2B	B2C
Foco exames?	Vestibulares	Não especifica	Não especifica
Pricing ENEM	R\$298,80	Depende das escolas	Não tem
Conteúdo educativo	Conteúdo padronizado	Atenção individualizada	Conteúdo padronizado

Número de usuários	1,5 milhão (total curso ENEM)	150 mil (total)	4 milhões / 200 mil pag (2021)
Fundação empresa	2011	2016	2012
Investimento	450 milhões (series E)	~R\$3 milhões	R\$27,7 milhões (seed, series A + B)

Fonte: autoria própria

Essa lista contém algumas empresas grandes e já bastante experientes com relação à educação online. Empresas como Descomplica e Passei Direto foram pioneiras no Brasil neste modelo de ensino, tendo começado com conteúdo educacional gratuito voltado para complementação escolar e encontrado tracção por conta de nosso profundo déficit educacional. São empresas que atendem justamente à grande demanda por acesso ao ensino superior e qualificação que abra oportunidades de melhora na qualidade de vida. A possibilidade de impactar dezenas de milhares de estudantes com aulas de qualidade gravadas e distribuídas na internet a um custo unitário muito baixo e surfando a onda de popularidade de plataformas como o YouTube fez do modelo de produção de vídeos e materiais educativos um muito bem sucedido neste contexto.

Na medida em que o acesso à internet se popularizou ao longo dos anos 2010, jovens com dificuldades na escola e em fase de preparação para vestibulares puderam recorrer a estas plataformas e ter nova chance de aprender conteúdos que passaram batidos. Essas 2 empresas têm como foco um ensino online baseado em materiais comprehensivos, voltados para a aprendizagem de conteúdos de modo autossuficiente, mas sem muito espaço para personalização. Estão inseridas no paradigma tecnológico que muitas vezes é chamado de “WEB 2.0”, que marca uma segunda onda de desenvolvimento de empresas na internet, não tanto orientadas por sites como na “WEB 1.0”, mas por plataformas sociais, e aplicativos a partir da ascensão do mobile.

Videoaulas, porém, não comprehendem todas as possibilidades da WEB 2.0, e isso é mostrado pelo modelo da TutorMundi. Em vez de produzir conteúdo educacionais e distribuir diretamente para estudantes que buscam recursos para aprender conteúdos, esta empresa criou uma plataforma em que conecta tutores (muitas vezes estudantes universitários) a alunos de escola, criando uma fonte de renda em formato flexível para profissionais e aliviando o gargalo

comum a muitas escolas de ter que suprir a demanda por dúvidas dos alunos com horas-aula escassas dos professores. Além disso, há a vantagem que os alunos podem tirar dúvidas de casa em qualquer momento. A inspiração principal da empresa é justamente a de oferecer uma solução ao problema Bloom-2-Sigma: ao conectar estudantes e tutores pela internet, ela previne ineficiências do modelo presencial como o tempo gasto no transporte e desajustes do modelo de remuneração “por hora” em relação à demanda, e torna a tutoria mais escalável. Esse modelo que já foi introduzido pensando na atenção individualizada deixou a TutorMundi muito bem posicionada para implementar IA Generativa em sua solução, o que vem fazendo desde 2024 com bastante efetividade: em apenas um trimestre, alcançaram 20 mil atendimentos nessa modalidade.

Em relação às plataformas EAD, fica muito clara a distinção dos modelos e soluções em relação à nossa proposta. Por terem como foco o conteúdo educacional viável em ampla escala, as empresas que atuam dessa maneira ao longo dos anos foram investindo pesadamente na produção e distribuição desse conteúdo para cobrir todo o currículo relevante. Há que se notar que em relação aos métodos, elas não diferem muito dos cursinhos presenciais, tendo os fundadores da Descomplica, por exemplo, tendo sido professores de cursinhos preparatórios (ENDEAVOR, 2019). Atualmente, a Descomplica está em um movimento de consolidar a oferta de cursos formativos EAD, de graduação e pós-graduação, e em razão disso receberam investimentos de centenas de milhões de reais. A Passei Direto se diferencia por ser uma plataforma aberta de compartilhamento de materiais de estudo e outros conteúdos, sem cursos delimitados e bem estruturados como as outras; foi adquirida pela UOL EdTech em 2021, em transação que não teve os valores anunciados (BRAZIL JOURNAL, 2021).

Consideramos a possibilidade de investimento público na criação de conteúdos digitais e EAD voltados ao reforço e revisão dos conteúdos compreendidos pelo currículo escolar, os quais são testados nos vestibulares, como uma estratégia altamente promissora. Esse tipo de investimento teria um enorme potencial de fornecer aos estudantes motivados, que não dispõem de recursos financeiros para pagar por plataformas como a Descomplica (R\$298,80) ou a PasseiDireto (R\$59,90), instrumentos de estudo flexíveis e acessíveis, provenientes de uma fonte oficial e construídos por especialistas. A proposta aqui apresentada, no entanto, em vez de focar na criação de novos conteúdos se concentra na adaptação dos conteúdos já disponíveis nos Modelos Fundacionais à individualidade de cada estudante, fornecendo uma tutoria adaptativa que otimize sua preparação para os exames. Essa distinção coloca as solução em um caminho distinto.

Já em relação à TutorMundi, há grandes semelhanças com a presente proposta, e vale fazer uma comparação mais direta destacando três distinções importantes. A primeira e mais evidente é em relação ao uso de tutores humanos via “*marketplace*” vs o uso nativo de IA. Em termos da estrutura do negócio, a monetização no *marketplace* se dá obrigatoriamente por comissionamento, assim como no caso de empresas como Uber, e há a necessidade de um direcionamento significativo de recursos na aquisição de usuários tutores e na gestão desses profissionais autônomos. Ao mesmo tempo, a empresa passa a ter que lidar com as questões de ética trabalhistas próprias da economia de plataformas.

Mas esse modelo também traz perspectivas muito interessantes em relação à implementação da IA para tutorias automatizadas, que a empresa já vem praticando. Em primeiro lugar, a TutorMundi pode testar e promover implementação gradual da tutoria baseada em LLMs na medida da aceitação dos alunos, contando com um marketplace validado e funcional como base. Ao longo do tempo, a substituição pela IA pode ser muito vantajosa para a empresa, que pode inclusive mudar a forma de atendimento com tutores para um formato que aproveite melhor a capacidade de formar uma conexão afetiva com os estudantes. Em segundo lugar, as tutorias com profissionais fornecem dados que podem ser usados no treinamento (*fine-tuning*) de LLMs, e isso pode ser a chave para uma adaptação bem sucedida de modelos fundacionais e da incorporação de maior consistência, é claro, respeitadas as questões legais de propriedade intelectual das interações que geraram esses dados. O *case* de implementação de IA voltado a tutoria individual pela TutorMundi é um para nos mantermos atentos, pois enfrentará alguns obstáculos similares aos nossos.

Uma segunda distinção interessante é que a TutorMundi opera com as escolas por contratos baseados em atendimento particular, sem ter material ou direcionamento próprio. A empresa oferece um modelo de tutoria em que o aluno toma a iniciativa de trazer suas dúvidas, para então os tutores humanos ou artificial agirem. Se de um lado eles possuem dados valiosos de treinamentos humanos e avaliações para treinar a IA, do outro lado não está sob seu controle a trilha de aprendizado dos estudantes, e não possuem facilidade de trabalhar os dados do ponto de vista dos conteúdos sob demanda. Ao utilizar um conjunto fixo de questões e manter o controle sobre a trilha de aprendizado, como é nosso caso, abre-se uma oportunidade nova de empregar inteligência analítica para trabalhar o progresso dos estudantes visando a otimização contínua.

A última diferença é em relação à missão e o modelo B2G vs B2B. A proposta da TutorMundi também visa viabilizar educação eficiente em ampla escala, sendo uma resposta para o problema Bloom-2-Sigma, mas ela não tem como objetivo direto a redução na

desigualdade de oportunidades. Ademais, o sistema de contratar tutores privados sob demanda dificilmente se ajustaria à educação pública, tendo a empresa foco total nas escolas privadas. Essa opção foi feita justamente porque ao ter escolas como clientes, agregam a vantagem que explicitamos na seção 3.1: cada venda para uma escola possui um impacto maior sobre a receita, ainda que o ticket médio por usuário possa ser baixo. Ao não ter o impacto social como objetivo primário, surge uma distinção importante que faz com que a TutorMundi não se desenvolva no mesmo sentido que faz sentido aqui.

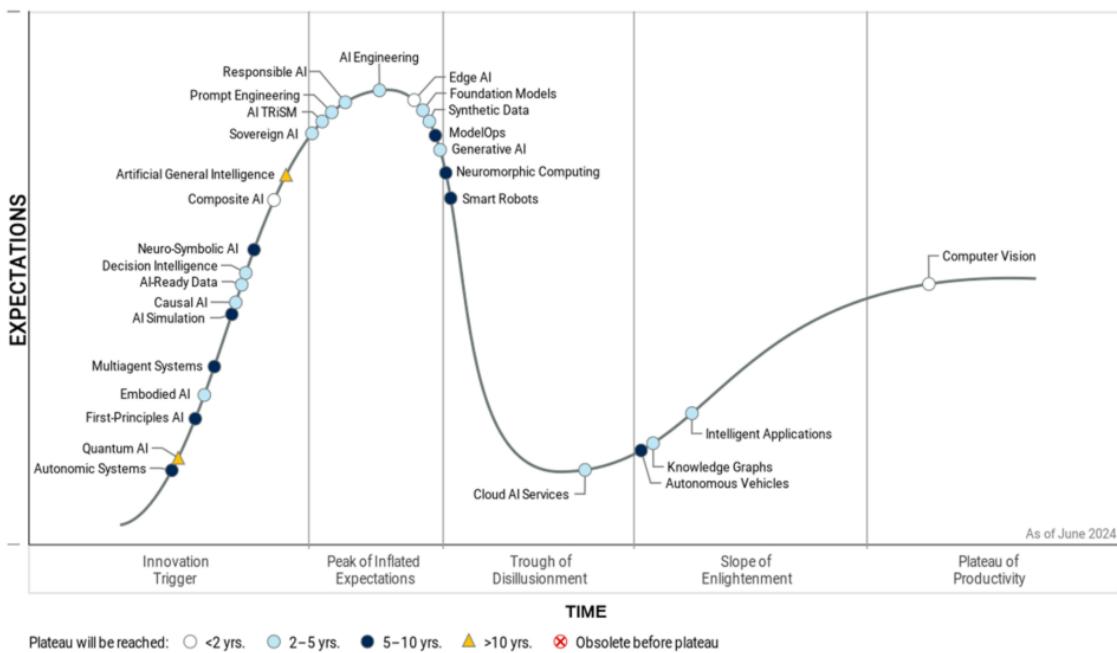
Foram indicados nessa seção algumas EdTechs como Descomplica, Passei Direto e TutorMundi que estão mudando o panorama da preparação para vestibulares no Brasil, e o avanço contínuo da inteligência artificial, especialmente com o desenvolvimento dos LLMs, promete transformar ainda mais o cenário educacional. O potencial dos LLMs de oferecer uma nova fronteira de inovação, com a capacidade de personalizar a experiência de aprendizagem de uma maneira que era inimaginável há poucos anos é amplamente reconhecida. Contudo, por se tratar de uma tecnologia emergente, as soluções baseadas em LLMs ainda estão em fase de desenvolvimento e experimentação. Na seção seguinte, o panorama atual do mercado de LLMs será apresentado, com uma análise de como essas tecnologias estão sendo aplicadas e quais oportunidades se apresentam para o desenvolvimento de uma solução educacional que não só se integre a essa nova onda tecnológica, mas também a utilize de forma pioneira para democratizar o acesso ao ensino de qualidade no Brasil.

3.3 Panorama de mercado de soluções educacionais baseadas em LLMs

3.3.1 A hype dos Large Language Models

Mais do que qualquer outro case, o ChatGPT marca o momento em que a inteligência artificial cativou a atenção do público e do mercado. O produto da empresa OpenAI foi lançado em novembro de 2022 para alcançar a marca de 100 milhões de usuários apenas 2 meses depois, a adesão mais rápida de qualquer produto em toda história (The Guardian, 2023), efetivando um período de ampla projeção dos LLMs nas fronteiras da aplicação de IA a negócios. Em relatório publicado em junho de 2024, o grupo Gartner entendia que a Inteligência Artificial Generativa (que abrange os LLMs) havia naquele momento ultrapassado o pico do ciclo de expectativas, com o desenvolvimento das grandes inovações relacionadas à tecnologia estando prevista para os próximos 2 a 5 anos, uma margem de tempo bastante curta.

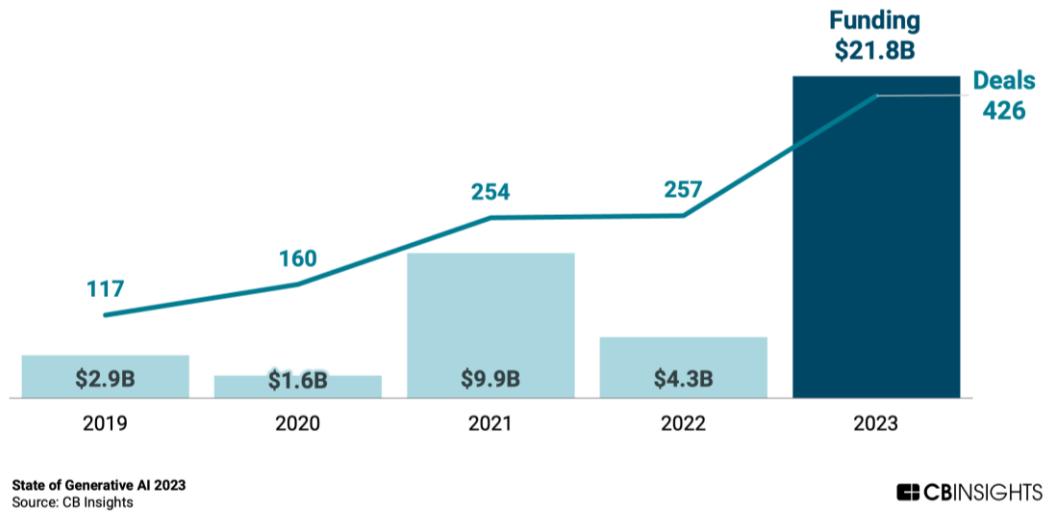
Figura 2 – Hype cycle for Artificial Intelligence, 2024



Fonte: Gartner.

A ‘hype’ sobre a IA Generativa tem muito a ver com há o efeito subjetivo de finalmente existir um computador capaz de conversação com pessoas e de simular raciocínios, novidade que ascendeu debates sobre a possibilidade de virem a ser desenvolvidas IAs com capacidade intelectual superiores aos de seres humanos. Também existe o lado sólido da empolgação, que é facilmente constatado quando considera-se com mais cuidado a importância do objeto de expertise dos LLMs. O fato de grande importância é o da aplicação bem sucedida de Redes Neurais Profundas para a língua, ou para o Processamento de Linguagem Natural, na terminologia da computação. A mesma tecnologia que quando bem aplicada em domínios específicos (Narrow AI) como os jogos de xadrez e Go ultrapassou a capacidade de todos os seres humanos, passou a operar com sucesso no amplo domínio da linguagem verbal (LECUN; BENGIO; HINTON, 2015). Dada a natureza pervasiva da linguagem verbal em relação a muitas das atividades produtivas humanas, e o envisionedo potencial dos LLMs de elevar a produtividade numa parcela significativa destas atividades, a expectativa quase consensual de analistas de instituições privadas e públicas de todo o mundo (STANFORD UNIVERSITY, 2023; MCKINSEY & COMPANY, 2023) é que ao longo dos próximos anos a tecnologia irá se imbricar ao cotidiano de uma população global que já passa em média 6 horas e 37 minutos diárias na internet (DATA REPORTAL, 2023) entre vida pessoal e trabalho.

Figura 3 – Investimentos globais em IA Generativa, 2019-2023



Fonte: State of Generative AI 2023, CBInsights.

Como mostra a figura 3, o momento é de grande otimismo e investimentos maciços na área, com um volume global de investimentos em IA Generativa tendo crescido 5x e atingido US\$ 21.8 bilhões em 2023, segundo a CBInsights (2024). Contudo, o Gartner (2024) já alertou para o risco de uma possível retração nos investimentos, caso as aplicações de LLMs não demonstrem resultados concretos e sustentáveis que justifiquem o capital empregado. Este cenário de incerteza reforça a necessidade de que as iniciativas baseadas em LLMs sejam cuidadosamente planejadas e fundamentadas em casos de uso robustos, que provem seu valor real na prática, e não apenas prometam retornos especulativos. Assim, o verdadeiro diferencial para empresas e investidores será a capacidade de identificar e desenvolver aplicações que não apenas acompanhem a tendência, mas que realmente integrem o valor agregado da IA generativa às operações e estratégias de negócio de forma mensurável e escalável.

Dado esse cenário de grande entusiasmo, mas também de incertezas, as oportunidades de captação de recursos via editais privados e públicos tendem a existir, mas reforçamos que essas possibilidades de financiamento exigem um planejamento estratégico cuidadoso e uma apresentação clara dos benefícios e diferenciais que as soluções propostas podem oferecer. A próxima seção explorará resultados de uma pesquisa sobre casos de implementações de LLMs em contextos educacionais reais, que hoje servem como nossas principais referências práticas e exemplos de viabilização.

3.3.2 Panorama de mercado de soluções educacionais baseadas em LLMs

Nos últimos anos, o mercado global de soluções educacionais baseadas em Large Language Models tem se expandido rapidamente, com várias plataformas emergindo para oferecer tutoria personalizada e suporte acadêmico em tempo real. Exemplos notáveis incluem o Khanmigo, da Khan Academy, que integra LLMs para fornecer uma experiência de aprendizado interativa e adaptativa em uma variedade de disciplinas (KHAN, 2024), e o Socratic, da Google, que utiliza inteligência artificial para auxiliar alunos na resolução de problemas complexos, oferecendo explicações detalhadas e recursos multimídia (GOOGLE, 2024). Outro exemplo relevante é o Riiid Tutor, uma plataforma sul-coreana que personaliza a experiência de estudo para exames padronizados através de análises preditivas (RIIID, 2024). Essas soluções demonstram o interesse em aplicar LLMs para revolucionar a tutoria educacional, criando ambientes de aprendizado mais acessíveis e personalizados; cada uma delas será contextualizada como exemplo valioso do que pode ser feito. São ferramentas ainda amplamente direcionadas a contextos internacionais, o que ressalta a importância de desenvolvermos uma solução específica para o contexto brasileiro, especialmente voltada para a preparação ao ENEM.

A Khan Academy, fundada em 2008 por Salman Khan, foi criada com o objetivo de oferecer uma educação gratuita e de alta qualidade para qualquer pessoa, em qualquer lugar do mundo. Inicialmente conhecida por suas videoaulas e exercícios interativos, a plataforma se consolidou como um recurso muito utilizado por estudantes ao redor do globo, impactando milhões ao fornecer acesso à conteúdos educacionais abrangentes e acessíveis (KHAN ACADEMY, 2023). Com uma missão centrada na democratização do conhecimento, a Khan Academy se tornou uma referência em educação online, alcançando um público diversificado e contribuindo para a redução das desigualdades educacionais em diversas partes do mundo.

Em 2023, a Khan Academy inovou ao lançar o recurso Khanmigo, um tutor virtual baseado em LLMs, desenvolvido em parceria com a OpenAI que buscava apresentar cases em educação de seu novo modelo, GPT-4. O Khanmigo opera dentro da interface da Khan Academy, ajudando os alunos a interagir de forma personalizada com o conteúdo, oferecendo explicações detalhadas, resolvendo dúvidas em tempo real e ajustando o ensino ao nível de compreensão de cada estudante. O uso de IA Generativa possibilita que conteúdos escaláveis como videoaulas e textos, inherentemente padronizados, possam ser adaptados para atender às necessidades distintas de seres humanos diferentes, sendo um exemplo magnífico do potencial da aplicação dos LLMs à educação. A solução ainda não está disponível em português. Um destaque deste caso de uso é a comprovação no valor da adaptação de conteúdos escaláveis para atender às

necessidades individuais dos alunos, o que valida a premissa de personalização da nossa solução. Por outro lado, é importante considerar que a Khan Academy integrou o Khanmigo como uma extensão de sua já consolidada plataforma de aprendizado, o que sugere que uma solução robusta pode ser construída sobre bases educacionais já estabelecidas; isso pode ser uma área de reflexão para nosso projeto já que ele depende da infraestrutura educacional pública.

Uma outra solução interessante a ser considerada por nós é o Socrative. O Socrative teve sua origem como um aplicativo independente, criado por um grupo de desenvolvedores com a missão de facilitar o aprendizado dos estudantes através da tecnologia. Em 2018, o Google adquiriu a plataforma, impulsionando seu desenvolvimento e integração com tecnologias de inteligência artificial mais avançadas (LYONS, 2019). O Socrative permite que estudantes façam perguntas por texto ou imagem e recebam explicações detalhadas, passo a passo. Suas principais funcionalidades incluem a capacidade de resolver equações matemáticas ao simplesmente tirar uma foto, fornecer resumos e explicações em linguagem simples, e recomendar recursos adicionais, como vídeos e artigos, para aprofundar o entendimento do estudante.

O Socrative destaca-se pela sua acessibilidade e interface intuitiva, demonstrando que a simplicidade e a facilidade de uso são fatores cruciais para o sucesso de soluções baseadas em LLMs. A ferramenta tem sido amplamente bem sucedida em oferecer assistência rápida e clara para dúvidas acadêmicas, mas vale notar que o Google não precisou fazer grandes ajustes em sua estratégia após a aquisição do Socrative, sugerindo que o modelo de perguntas-respostas já estava bem alinhado às necessidades dos usuários. Isso nos ensina que, para nossa solução, é importante manter uma interface amigável e focada na simplicidade, especialmente ao lidar com um público diverso de estudantes que pode incluir aqueles com menos familiaridade com tecnologias avançadas.

Outro case internacional, que nos interessa especialmente é o da Riiid, uma empresa sul-coreana que tem se destacado no campo da educação com um forte compromisso com o impacto social. Fundada em 2014, a Riiid começou com a missão de democratizar o acesso à educação de alta qualidade através da inteligência artificial. A empresa rapidamente ganhou notoriedade ao desenvolver soluções de aprendizado que utilizam tecnologias avançadas para personalizar a experiência educacional de cada aluno, independentemente de suas origens socioeconômicas. Além do impacto social, a empresa também atua com foco em exames padronizados, como o TOEIC, sendo um exemplo de sucesso por ter conseguido impactar um grande número de

estudantes que buscam melhorar suas habilidades e competir em pé de igualdade em um mercado de trabalho global cada vez mais competitivo (CHUNG, 2022).

O sistema da Riiid, impulsionado por LLMs, oferece uma série de features avançadas que tornam a aprendizagem mais eficaz e adaptativa. O Riiid Tutor, uma das principais plataformas da empresa, utiliza análise preditiva para identificar pontos fortes e fracos dos alunos, ajustando automaticamente o conteúdo e a dificuldade dos exercícios com base no desempenho individual. A abordagem da Riiid em combinar LLMs com análise preditiva para personalizar o aprendizado é altamente relevante para nosso projeto, sugerindo que o uso de análise de dados pode ser um diferencial competitivo importante. A Riiid Tutor tem sido bem sucedida, especialmente por seu foco em exames padronizados, um segmento que também é central para a nossa proposta. Isso valida a viabilidade de nossa solução, mas também destaca a importância de continuamente refinar nossos algoritmos para garantir que o feedback fornecido aos alunos seja o mais preciso e útil possível.

Por fim, um exemplo brasileiro que merece destaque é a Cubos Academy, uma EdTech que se dedica a capacitar profissionais para o mercado de tecnologia. Entre seus produtos, a Emy se destaca como uma solução inovadora inspirada em ferramentas globais como o Khanmigo. A Emy é um tutor virtual alimentado por LLMs, desenvolvido para auxiliar os alunos da Cubos Academy durante os cursos de programação e outras disciplinas técnicas. Inspirada no sucesso de plataformas como o Khanmigo, a Emy foi criada para oferecer suporte personalizado em tempo real, respondendo às dúvidas dos alunos, guiando-os através de exercícios práticos e adaptando-se ao nível de conhecimento de cada usuário (FAST COMPANY, 2023). A partir do sucesso, a empresa começou a ofertar a integração da Emy para outras plataformas educacionais baseadas em conteúdo, como a Descomplica, tornando a Cubos Academy uma das pioneiras no Brasil a utilizar essa tecnologia em larga escala para a educação técnica e impactando milhares de estudantes no país.

A Emy, da Cubos Academy, oferece insights valiosos para nossa solução, especialmente no contexto brasileiro. A adoção de um tutor virtual baseado em LLMs em um ambiente educacional técnico demonstrou que há uma demanda significativa por suporte adaptativo em tempo real, o que fortalece nossa convicção de que essa abordagem será igualmente bem recebida no contexto de preparação para o ENEM. Além disso, o sucesso da Emy na integração com outras plataformas, como a Descomplica, sugere que parcerias estratégicas podem ser um caminho promissor para a escalabilidade e o impacto do nosso projeto.

3.4 Análise estratégica

Para fechar essa seção do trabalho, é apresentada uma análise estratégica que incorpora aprendizados da pesquisa de modelos e de mercado. Para isso, são utilizadas duas ferramentas amplamente reconhecidas: a matriz SWOT e o Business Model Canvas. Essas ferramentas permitem avaliar de forma holística os fatores internos e externos que influenciam o presente projeto e delinear um modelo de negócios eficiente e adaptado ao seu contexto.

A matriz SWOT (*Strengths, Weaknesses, Opportunities, Threats*) é uma ferramenta de planejamento estratégico criada na década de 1960 por Albert Humphrey, enquanto trabalhava no Instituto de Pesquisa de Stanford (NIH, 2018). Ela é utilizada para identificar e analisar as forças e fraquezas internas de um projeto ou organização, bem como as oportunidades e ameaças externas. Ao aplicar a matriz SWOT (tabela 4) ao presente caso, podemos entender melhor onde a solução proposta se destaca, quais são os desafios internos que precisam ser superados, quais oportunidades de mercado estão disponíveis e quais ameaças externas podem impactar seu sucesso. Essa análise é crucial para ajustar a estratégia e maximizar as chances de sucesso do projeto.

Tabela 4 – Matriz SWOT da solução.

Strengths	Opportunities
<ul style="list-style-type: none"> • Tecnologia inovadora “no core”: solução alavanca as principais características de uma tecnologia poderosa e inovadora • Custos <i>lean</i>: estrutura de custos extremamente simples e enxuta • Desenvolvimento inicial facilitado: flexibilidade de LLMs de ponta e o direcionamento oferecido pelos conteúdo do ENEM • Acessibilidade e de impacto social: potencial de entregar muito valor ao 	<ul style="list-style-type: none"> • Facilidade de desenvolvimento de MVP: ferramentas disponíveis possibilitam produzir um MVP de baixo custo e melhorar iterativamente a solução • Demanda crescente por educação personalizada: crescente interesse por soluções de aprendizado que se adaptem às necessidades individuais dos estudantes • Expansão para outros exames e mercados: possibilidade de

<p>capacitar para acesso ao ensino superior com baixo custo</p> <ul style="list-style-type: none"> • Parcerias estratégicas: possibilidade de colaboração com escolas públicas e privadas, além de parcerias com fundações e instituições filantrópicas que buscam apoiar a educação no Brasil 	<p>adaptar a solução para outros exames importantes</p> <ul style="list-style-type: none"> • Tendência de IA na educação: momento favorável para captação de investimentos e parcerias, aproveitando o crescente interesse em inteligência artificial aplicada à educação
<p>Weaknesses</p> <ul style="list-style-type: none"> • Modelo não provado: criar solução eficaz e engajante é grande desafio, e alucinações oferecem riscos • Recursos limitados: necessidade de financiamento significativo para o desenvolvimento, testes e aprimoramento contínuo da plataforma • Dependência de infraestrutura tecnológica: necessidade de conectividade de internet de alta qualidade para que os estudantes possam acessar o tutor virtual de forma eficaz; dependência de APIs de empresas proprietárias de modelos fundamentais • Falta de conhecimento do público: uso de IA e instrução automatizada pode gerar desconfiança • Necessidade de escala: preciso alta escala para viabilizar a solução e ter base significativa para uso estratégico de dados 	<p>Threats</p> <ul style="list-style-type: none"> • Limitação nas opções de financiamento público: poucas possibilidades ameaçam a realização total do projeto • Concorrência no uso de modelos fundamentais: falta de diferenciação na implementação de LLMs pode criar muitos competidores e fazer uma presa fácil para concorrentes consolidados • Desafios na escalabilidade: dificuldade em escalar a solução para atender a um grande número de estudantes de forma eficaz e sustentável • Riscos de regulamentação: potenciais mudanças na regulamentação educacional e em privacidade de dados que poderiam restringir o uso de tecnologias de IA em escolas

<ul style="list-style-type: none"> • Dependência de base educacional: a plataforma poderá sanar dúvidas, mas não ensinar os conteúdos, portanto depende de uma base e pode ser insuficiente para promover preparação ao ENEM 	<ul style="list-style-type: none"> • Falta de esperanças no vestibular: muitos estudantes de escolas públicas não têm esperanças de cursar ensino superior, o que pode afetar disposição a usar nossa solução
---	--

Fonte: elaboração própria.

A análise da matriz SWOT revela um panorama estratégico que confirma o alto potencial do projeto ao mesmo tempo que identifica desafios significativos que precisam ser superados para garantir o sucesso a longo prazo.

Do lado das forças, a solução proposta se beneficia de uma tecnologia altamente inovadora e poderosa em seu núcleo, e da adequação a um contexto em que soluções educacionais de baixo custo são muito demandadas. Ao alavancar LLMs de ponta, a proposta tem o potencial de transformar a preparação para o vestibular, adaptando-se ao nível de compreensão e às necessidades específicas de cada aluno. Essa capacidade de personalização responde diretamente à crescente demanda por educação individualizada, posicionando-nos de maneira interessante em relação às principais tendências de uso de tecnologia para fins educacionais e de instrução.

Além disso, a estrutura de custos *lean* e a facilidade de desenvolvimento inicial permitem que o projeto avance de forma ágil, com menos dependência de grandes aportes financeiros no início. Essa abordagem enxuta é essencial em um cenário de recursos limitados, permitindo iterar e refinar a solução com base no feedback dos usuários. Outro ponto de destaque é o potencial de impacto social do projeto, que visa democratizar o acesso ao ensino superior, especialmente para estudantes de contextos socioeconômicos desfavoráveis. Esse aspecto pode atrair parcerias estratégicas e financiamento de entidades filantrópicas e do setor público, que buscam reduzir desigualdades educacionais no Brasil.

No entanto, a análise também revela desafios significativos que precisam ser abordados. Um dos principais riscos é a necessidade de provar a eficácia e o engajamento da solução junto aos usuários. O fenômeno das "alucinações" dos modelos de IA, que podem gerar informações incorretas ou irrelevantes, representa uma ameaça à confiança dos usuários e à credibilidade da plataforma. Portanto, é crucial implementar estratégias robustas de validação e melhoria contínua do modelo, começando com uma estratégia de MVP muito bem pensada. A

necessidade de escala para viabilizar a solução e captar dados de forma estratégica é um desafio que depende da capacidade de progredir no desenvolvimento da solução com eficiência.

A dependência de conectividade de internet de alta qualidade e de APIs de empresas proprietárias de modelos fundacionais acrescenta riscos operacionais e financeiros que precisam ser geridos cuidadosamente, mas o alto ritmo de inovações tanto nos modelos fundacionais quanto na infraestrutura tecnológica, como o avanço do 5g, nos são favoráveis. Outro fator que merece atenção é a intensa concorrência no uso de modelos fundacionais, especialmente se grandes players globais decidirem entrar no mercado brasileiro com soluções semelhantes. Isso aumenta a pressão para diferenciar nossa proposta de valor de maneira clara e eficaz. A limitação de fontes de financiamento público também pode restringir o alcance e a expansão do projeto, exigindo uma busca proativa por alternativas de financiamento privado ou parcerias estratégicas que possam fornecer os recursos necessários.

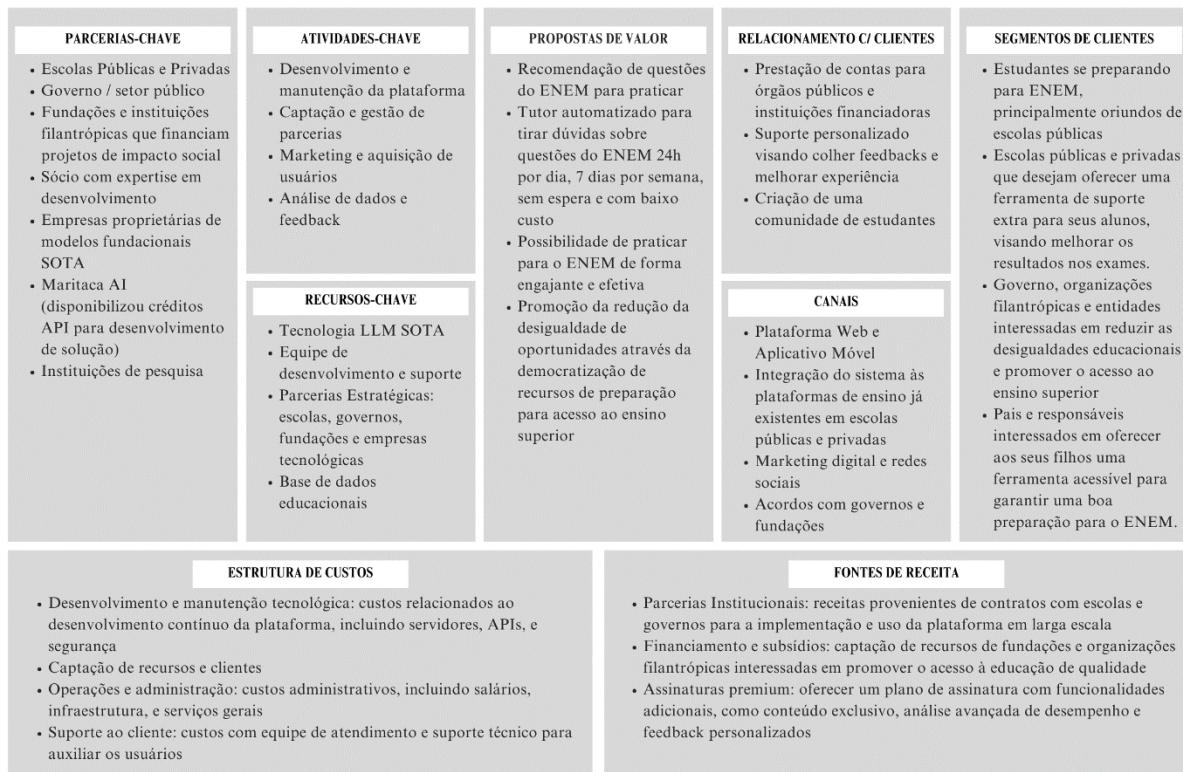
Para aumentar as chances de sucesso do projeto, é proposta uma estratégia *go-to-market* com alguns elementos básicos.

1. Desenho e realização de experimentos muito bem estruturados que vão no sentido da validação das principais hipóteses da solução e da mitigação dos riscos relacionados ao projeto;
2. Especificação de um MVP a ser desenvolvido com uma estrutura de custos enxuta voltado, além da validação da entrega de valor, para a diferenciação em relação às empresas incumbentes;
3. A elaboração de um *roadmap* (plano de desenvolvimento da solução para além do MVP) e um cronograma iterativo de desenvolvimento, etapa por etapa, começando com o lançamento da versão mais simples possível e da colheita de feedbacks de usuários;
4. O início rápido do desenvolvimento da solução a partir do uso dos recursos disponíveis e a atualizando constante do projeto com os aprendizados;
5. O monitoramento constante de editais relacionados a inovações na educação e o uso dos resultados obtidos para aumentar as chances de obter financiamento e parcerias estratégicas;

Outra ferramenta muito útil para orientar o desenvolvimento do projeto é o *Business Canvas*. Desenvolvido por Alexander Osterwalder e Yves Pigneur, é uma ferramenta visual que permite esboçar e iterar rapidamente diferentes modelos de negócios (INTERACTION DESIGN, 2023). Ele é composto por nove blocos que descrevem os principais aspectos de qualquer negócio: Proposta de Valor, Segmentos de Clientes, Canais, Relacionamento com

Clientes, Fontes de Receita, Recursos-Chave, Atividades-Chave, Parcerias-Chave e Estrutura de Custos. Ao aplicar o Canvas a este projeto, pode-se mapear de forma clara e concisa como nosso produto criará valor, como será entregue aos clientes, e como gerará receita. Essa ferramenta permite testar diferentes cenários e refinar nosso modelo de negócios para garantir que seja viável e sustentável no longo prazo, além de alinhar nossa estratégia com as necessidades do mercado e as capacidades da nossa organização.

Figura 5 – Business Canvas de projeto de tutoria para ENEM com IA Generativa



Fonte: elaboração própria.

A análise por meio do Canva permite integrar os principais componentes essenciais para a viabilidade do projeto e a promoção do impacto social pretendido. Todos os componentes visam servir a proposta de valor central do projeto, de oferecer uma plataforma de tutoria acessível e personalizada voltada a preparação ao ENEM, utilizando modelos de linguagem avançados. Tendo em vista o objetivo de além de proporcionar uma experiência educacional adaptativa e interativa, oferecer recursos acessíveis de preparação ao vestibular no sentido de democratizar o acesso ao ensino superior e oportunidades, é preciso conciliar elementos de empresas inovadoras com os de projetos sociais.

Os segmentos de clientes definidos naturalmente incluem estudantes em preparação para o vestibular, sobretudo no ensino médio público que se preparam para o ENEM e escolas

públicas e privadas que buscam melhorar os resultados de seus alunos. Mas antes é necessário adequar o projeto aos editais promovidos pelo governo e organizações filantrópicas interessadas em reduzir as desigualdades educacionais, bem como desenvolver processos de prestação de contas eficientes.

A plataforma estará disponível tanto em dispositivos móveis quanto em computadores, facilitando o uso em diversos contextos e permitindo que a tutoria chegue a um público amplo. Parcerias com escolas e governos são componentes críticos nesse cenário, pois não apenas funcionam como canais de distribuição, mas também ajudam a integrar a solução ao sistema educacional existente, ampliando sua escalabilidade.

No que diz respeito às fontes de receita, o modelo traz como principal fonte as parcerias institucionais com escolas e governos que podem gerar receitas significativas através de contratos para a implementação e uso da plataforma em larga escala. A captação de financiamento e subsídios junto à fundações e organizações filantrópicas que compartilhem o objetivo de promover o acesso à educação de qualidade é outra opção importante. O Edital da Fundação Itaú “IA para Educação”, apresentado no Anexo A, estabelece prêmios de até R\$ 200.000 para desenvolvimento inicial de soluções como a apresentada aqui, um valor que é suficiente tendo em vista o orçamento apresentado na seção 4. Até chegar a uma parceria definitiva, poderá depender de assinaturas individuais para cobrir os custos de desenvolvimento e API ou em contratos junto a escolas privadas.

Os recursos-chave para a operação desse modelo incluem a tecnologia de LLMs, que é o núcleo da plataforma, além de uma equipe qualificada de desenvolvimento e suporte, capaz de manter e aprimorar continuamente o sistema. As parcerias estratégicas, tanto com instituições públicas quanto com empresas tecnológicas, são essenciais para viabilizar a implementação e expansão da solução. A base de dados educacionais, alinhada ao currículo do ENEM, é outro recurso crucial, garantindo que o conteúdo oferecido seja relevante e adaptado às necessidades dos alunos.

As atividades-chave necessárias para a execução do modelo incluem o desenvolvimento contínuo da plataforma, a captação e gestão de parcerias estratégicas, a execução de eventuais campanhas de marketing para aquisição de usuários, e a análise de dados para o aprimoramento contínuo da tutoria personalizada. Essas atividades são sustentadas por uma estrutura de custos que inclui despesas com desenvolvimento tecnológico, marketing, suporte ao cliente, operações administrativas e licenciamento de tecnologia.

Por fim, as parcerias-chave com escolas, empresas de tecnologia, governos e organizações filantrópicas desempenham um papel central no modelo. Essas parcerias não

apenas viabilizam o desenvolvimento e a implementação da solução, mas também ampliam seu alcance e impacto, conectando a plataforma a um público mais amplo e diversificado.

4 MVP: ESTRUTURAÇÃO E PLANEJAMENTO

4.1 Entendendo o papel de um MVP

A ideia de *Minimum Viable Product* (MVP) foi introduzida primeiramente pelo empreendedor e autor Eric Ries (2012), como parte de sua metodologia denominada *Lean Startup*, ou startup enxuta. A principal finalidade da metodologia é a de reduzir riscos e prevenir desperdícios de recursos em um contexto de empreitadas comerciais inovadoras em que a principal característica é a incerteza, aumentando assim suas chances de sucesso. Para isso, o princípio mestre dessa metodologia é o da experimentação: deve-se testar hipóteses relativas à solução e produto elaborados junto ao mercado com usuários reais assim que possível, colhendo informações importantes para a continuidade e evitando ao máximo investir recursos em aspectos não comprovados da solução, que podem acabar implicando em desperdício de recursos escassos.

O MVP é um componente essencial dessa metodologia, sendo importante entender bem o que significa e qual o seu papel. O MVP é a “primeira versão” de um produto a ser testado no mercado, possuindo dois componentes principais. *Viable*, ou viável, se refere ao aspecto de que essa versão do produto deve ser capaz de entregar o valor pretendido na proposta de valor. *Minimum*, ou mínimo, se refere ao aspecto da versão do produto de que ela deve ser a versão mais simples e menos exigente em termos de recursos necessária para entregar o valor pretendido. Uma boa forma de se compreender o MVP é conceber uma especificação de todas as possibilidades de desenvolvimento de um produto que atende a uma dada proposta de valor. Destes, deve se selecionar a mais simples de todas, e este será o MVP.

Existem inúmeros bons exemplos de MVPs que acabaram desempenhando um papel importante na criação de soluções inovadoras por startups bem sucedidas. Um caso famoso brasileiro que representa bem o conceito é o da Easy Taxi, empresa criada em 2011 propondo uma solução digital para agendamento de viagens de táxi. A viabilidade técnica da solução já era conhecida, mas era necessário verificar a inclinação do público alvo a utilizá-la. A tese elaborada que justificava a empresa pode ser compreendida por meio de quatro hipóteses: (1) um número significativo de passageiros enfrentam o problema de precisar gastar tempo e energia para achar táxis, esperando na rua ou pesquisando por um ponto e fazendo uma ligação, e lhes seria preferível solicitar através da internet pelo celular; (2) um número significativo de taxistas enfrentam o problema de precisar gastar tempo, energia e combustível para achar passageiros, e lhes seria preferível receber chamadas através da internet pelo celular; (3) pelo

menos um destes públicos está disposto a pagar para resolver seu problema, viabilizando o desenvolvimento e operação da solução.

O MVP executado foi bastante engenhoso: com recursos mínimos, foi criada uma página online que recebia solicitações de táxis com informações de contato do passageiro, endereço de partida e endereço de chegada. A página foi divulgada por meio de mídia paga na internet, e cada solicitação gerava o envio de um e-mail para um membro da equipe responsável, que obtinha informações no Google Maps de pontos de táxi mais próximos do passageiro e ligava para agendar uma passagem. Confirmações eram feitas com o passageiro e motorista, e mais uma quando o passageiro embarcava, e isso era controlado numa planilha (CAROLI, 2023). Um *co-founder* da empresa reporta que por meio deste exercício pôde-se comprovar a demanda por uma solução como essa do lado dos consumidores, que isso estimulou a equipe a seguir com o desenvolvimento e que foi o princípio da trajetória de uma startup bem sucedida.

Há três aspectos que valem ser destacados com relação a este caso. O primeiro e mais evidente é o da extrema simplicidade da solução elaborada, mas que ainda assim era suficiente para suprir um aspecto chave da proposta de valor. O segundo é que este MVP não permitiu validar todas as hipóteses necessárias para implicar viabilidade da solução ideal. Este experimento permitiu validar a hipótese 1, mas não a 2 e a 3. Do ponto de vista dos taxistas, o MVP não significou uma diferença, já que seguiram recebendo chamadas telefônicas no ponto. Ao mesmo tempo, esse serviço inicial foi oferecido de forma gratuita, não podendo determinar se havia disposição a pagar por passageiros ou motoristas. Para a continuidade da Easy Taxi, foram necessárias outras versões em crescente complexidade para validar as hipóteses principais, processo que diminuiu os riscos para a equipe e posteriormente para investidores. O terceiro, é que não basta desenvolver um MVP, mas é necessário levá-lo ao público correto que vê valor nele; caso contrário, não será possível validar o interesse na solução.

Pode-se perceber que há algo de científico no processo de elaboração e teste de um MVP, que é importante ter hipóteses bem estabelecidas, uma escala de prioridade, e procedimentos definidos para avaliação do experimento. Ao mesmo tempo, não é de maneira alguma um processo exato, havendo muito espaço para variação e uso de criatividade.

4.2 Formulação da tese e hipóteses

A tese que justifica o modelo de negócios aqui proposto pode ser apresentada da seguinte maneira:

Modelos fundacionais (LLMs) podem ser adaptados para se tornar uma ferramenta efetiva e engajante na preparação de estudantes do ensino público ao ENEM, e tal solução é viável do ponto de vista econômico.

Essa tese pode ser separada nas seguintes hipóteses, elencadas em ordem de prioridade:

1. LLMs *state-of-the-art* (do estado da arte, ou SOTA) são capazes de elaborar explicações que abordam as questões do ENEM sob diferentes ângulos e formatos com bom nível de precisão;
2. Estudantes em preparação ao ENEM sofrem da falta de opções de baixo custo para tirar dúvidas sobre a prova e há interesse em contar com uma ferramenta automatizada baseada em IA que atenda a essa necessidade;
3. O uso continuado dessa ferramenta pode aumentar o desempenho dos usuários oriundos do ensino público no ENEM;
4. Há interesse de governos ou escolas ou estudantes individuais suficiente para viabilizar a solução do ponto de vista econômico;

A partir dessas hipóteses, podem ser desenhados experimentos para validá-las. A solução proposta aqui depende da aplicação de uma tecnologia relativamente nova, por isso existe um componente técnico a ser testado. Um papel essencial de um experimento, neste caso, é o de avaliar a capacidade dos LLMs *state-of-the-art* não apenas para resolver questões do ENEM (que já foi verificada), mas para produzir explicações úteis para atender a questões de diferentes tipos. Para validar a hipótese 1, não é possível evitar o uso de LLMs da mesma maneira com que a Easy Taxi evitou o desenvolvimento de um *back-end* para conectar passageiros a taxistas.

Depois, há a necessidade de avaliar a demanda e percepção do público. Para isso, é necessário levar a proposta inicial de solução a usuários potenciais, e colher métricas ou *feedbacks*. Esse é um experimento com finalidade totalmente distinta do destinado a validar a hipótese 1, não sendo necessário utilizarem um mesmo MVP (pode ser ou não ser o mesmo). Para ilustrar isso, considere-se o seguinte caso: afim de validar a demanda com relação a uma determinada solução, é desenvolvida uma versão inicial desta solução, que é levada a um público e a demanda é avaliada a partir da análise de *feedbacks* coletados. Mesmo que os *feedbacks* sejam predominantemente negativos, isso não significa que não existe demanda para solução; pode ser que, por exemplo, o sistema desenvolvido não seja bom o bastante. Essa dependência entre as hipóteses deve ser evitada para a estruturação dos experimentos. Caso

opte-se por usar o mesmo MVP para validar as hipóteses 1 e 2, é fundamental ter clareza em relação às métricas de validação de cada uma.

A hipótese sobre a efetividade da solução na preparação ao ENEM é muito mais complexa, pois depende tanto de um produto mais robusto e de uma amostra de usuários relevante. Tendo em vista o conceito de MVP, uma vez que a proposta de valor depende essencialmente da efetividade da ferramenta na preparação ao ENEM, essa hipótese é uma central a ser validada. Isso não será possível com um experimento simples, sendo necessário elaborar um MVP bem mais complexo do que o necessário para testar a hipótese 1, e isso exigirá recursos adicionais.

Já em relação à viabilidade econômica, ela consiste em um caso totalmente diferente. Aqui realmente não faz sentido atrelar uma versão inicial de produto a uma monetização com expectativas de que isso apresente uma relação econômica próxima do projeto concreto. Faz mais sentido avaliar a viabilidade de diferentes modelos de monetização através de pesquisas exploratórias e experimentais, e mesmo efetuando buscas por editais com potencial de financiar o projeto.

4.3 Definição de experimentos preliminares

Serão apresentados a seguir experimentos detalhados para validar cada uma das hipóteses a serem testadas.

4.3.1 Hipótese 1 (técnica)

Para validar a hipótese “LLMs state-of-the-art são capazes de elaborar explicações que abordam as questões do ENEM sob diferentes ângulos e formatos com bom nível de precisão” é necessário desenvolver um protótipo que, tendo em vista a discussão feita nas seções anteriores, sobretudo a teórica, deve atender aos seguintes requisitos básicos:

1. Base de dados com questões históricas do ENEM contendo enunciados, alternativas e gabarito;
2. Um programa que efetue solicitações de API do modelo *State-of-the-Art* selecionado conectadas a um sistema de prompts efetivos;
3. Sistema RAG ou semelhante pelo qual o modelo acesse questões da base e seja capaz de reproduzi-las com fidelidade;
4. Algum programa recomendador de questões que oriente a seleção de questões da base;

5. Interface de *Chatbot* que permita interação entre modelo e usuário, em que este visualize questões, tire dúvidas e receba instruções da IA;

Destes, o requisito 2 é o principal, pois é dele que depende o sucesso do MVP na validação da hipótese sugerida. Nunca foi a intenção desenvolver um LLM proprietário, nem mesmo partindo de um modelo *open-source*, por isso conta-se com as capacidades e conhecimentos inerentes aos modelos SOTA para interpretar corretamente o enunciado de questões do ENEM e sua relação com cada uma das alternativas. Assim, um ponto crítico a ser validado é o desempenho dos modelos por meio de diferentes técnicas de prompt.

A partir dos resultados empíricos trazidos na seção 2, que demonstram a assertividade dos modelos GPT-3.5, GPT-4 e Sabiá-2, o próximo passo é testar esses modelos em relação às suas capacidade de administrar questões fornecidas por meio de prompts numa estrutura de chat e de ajudar estudantes a superar dificuldades e responder a questão corretamente.

Para isso, foram criados prompts que traziam exemplos de questões satisfazendo o limite da *context-window* (o limite de palavras que o modelo suporta dentro do prompt) e uma sequência de passos básicos que o sistema deveria seguir para simular a interação com um usuário buscando ajuda para se preparar ao ENEM. Esses passos são:

- 1) Selecionar aleatoriamente uma questão dentre as oferecidas no prompt;
- 2) Reproduzir a questão fornecida para o usuário, que deveria tentar responder em seguida;
- 3) Verificar a resposta correta no gabarito e retornar uma resposta ao usuário. Em caso de acerto, parabenizar o usuário e perguntar se gostaria de seguir praticando. Em caso de erro, comunicar o erro e oferecer uma explicação que não indicasse a alternativa correta, mas que tentasse estimular a compreensão do usuário.
- 4) Repetir passo 2 até acerto do usuário.

Figura 5 – Criação de função para *Chatbot* a partir de API da OpenAI

```
[3] # criando uma função para chatbot que é inicializado com um prompt, responde até usuário inputar "parar" e retorna o histórico de mensagens que pode ser usado para inicializar outro chat
def chatbot(prompt, model="gpt-3.5-turbo", temperature=0):
    history = [{"role": "system", "content": prompt}]
    print("Bem vindo a seu tutor para o ENEM. Peça uma questão para começar a praticar, e para finalizar digite 'parar'")
    user_prompt = "start"
    while user_prompt.lower() != "parar":
        user_prompt = str(input())
        history.append({"role": "user", "content": user_prompt})
        response = client.chat.completions.create(
            messages = history,
            model = model,
            temperature = temperature
        )
        assistant_response = response.choices[0].message.content
        history.append({"role": "assistant", "content": assistant_response})
        print("\nResposta do Assistente: ", assistant_response, "\n")
    return history
```

Fonte: elaboração própria.

Mesmo com um prompt simples *zero-shot* (sem utilização do *In-context-learning*), o modelo GPT3.5 foi capaz de seguir os passos estipulados, comprovando a destreza do modelo em seguir instruções. Não foi capaz de fazer isso todas as vezes, porém, havendo alguma margem de erro. Os principais erros do modelo se deram ao verificar a resposta correta no gabarito da questão selecionada e também na forma de instruir o usuário sem oferecer a alternativa correta.

Figura 7 – Testes iniciais com API da OpenAI no contexto de tutoria para prática direcionada do ENEM

```
v Iniciando testes para tutor do ENEM

[7] response = client.chat.completions.create(
    model = "gpt-3.5-turbo",
    messages = [
        {"role": "system", "content": "prompt"},
        {"role": "user", "content": "Olá, me indique uma questão para eu praticar"}
    ]
)
print(response.choices[0].message.content)
# bem razoável

[8] Mostrar saída ocula

[8] # one shot prompt mais simples possível (balizador)
prompt = f"**Você é um tutor que ajuda estudantes brasileiros na preparação ao ENEM. Para isso, a cada sessão selecione de forma aleatória uma questão da lista contida por 3 crises e a apresente ao usuário sem revelar o gabarito. Depois que ele fizer uma tentativa, verifique o gabarito correspondente na lista entre 3 crises; se a alternativa escolhida pelo usuário estiver errada explique o porquê da estar errada (IMPORTANTE: SEM DAR PISTAS DA RESPOSTA CERTA), e pergunte se o usuário deseja fazer uma nova tentativa; repita até que ele acerte. Se ele acertar, pergunte se ele deseja praticar com uma nova questão, retornando no passo anterior."
``[lista_questões]``"
````[lista_questões]````

[8] # testando chatbot com one-shot
teste_1 = chatbot(prompt=prompt)
a partir deste prompt, o modelo foi capaz de se comportar dentro do esperado: destacadamente selecionou uma questão e formatou adequadamente, inclusive removendo o gabarito
também foi capaz de não dar pistas das respostas. Dificuldades: identificar a alternativa correta, e oferecer explicações úteis
dado que o objetivo do sistema é ser um tutor, as explicações precisam ir além da formulação textual para explicar o que invalida respostas possíveis

[8] Ben vindo a seu tutor para o ENEM. Peça uma questão para começar a praticar, e para finalizar digite 'parar'.
Me indique uma questão

Resposta do Assistente: Claro! Aqui está a questão:

QUESTÃO 88 (gabarito: C)

De quem é esta língua?
Uma pequena editora brasileira, a Urutau, acaba de lançar em Lisboa uma "antologia antirracista de poetas estrangeiros em Portugal", com o título Volta para a tua terra. O livro reúne poemas de autores que, segundo o prefácio, os imigrantes estão sujeitos. Alguns dos poetas brasileiros antropólogos queixam-se do desdém com que um grande número de portugueses acolhe o português brasileiro.
É um tema frequente.
"Apesar de Portugal elas dizem / - elas dizem / - que nosso português é errado, que nós não falamos português", escreve a poetisa paulista Maria Cláudia Pinheiro, para concluir: "Se tua língua é tua bixinha, / ainda conserva a palerma da opressão / ela não é a sali bonita do mundo./ Ela é uma das mais violentas".
AGULUSA, J. E. Disponível em: http://ogolinhoaberto.com.br/2018/07/05/uma-litania-de-angustia/.
O texto de Agulusa tentava e preconceito em relação ao português brasileiro. Com base no trecho citado pelo autor, infere-se que esse preconceito se deve
A dificuldade de consolidação da literatura brasileira em outros países.
B aos diferentes graus de instrução formal entre os falantes de língua portuguesa.
C à existência de uma língua ideal que alguns falantes lusófonos acham superior ao português.
D ao intercâmbio cultural que ocorre entre os povos dos diferentes países de língua portuguesa.
E à distância territorial entre os falantes do português que vivem em Portugal e no Brasil.

Qual alternativa você escolhe e por quê?

A
Resposta do Assistente: A alternativa A está incorreta. O preconceito em relação ao português brasileiro, conforme apresentado no texto, não está relacionado à dificuldade de consolidação da literatura brasileira em outros países. O preconceito
C
Resposta do Assistente: Parabéns! A alternativa C está correta!
O preconceito em relação ao português brasileiro, conforme apresentado no texto, se deve à crença de alguns falantes lusitanos de que a língua falada em Portugal é a ideal, desqualificando o português brasileiro. Esta ideia de superioridade lingüística
Você gostaria de praticar com mais uma questão ou tem alguma dúvida?
parar
Resposta do Assistente: Certo! Se precisar de ajuda no futuro ou quiser praticar mais questões, estarei por aqui. Tenha um ótimo dia!
```

Fonte: elaboração própria.

O passo seguinte foi testar diferentes abordagens de prompting, incluindo *one-shot*, *few-shot* e *Chain-of-Thought* no mesmo modelo GPT-3.5, e o que pôde ser observado é que essas técnicas não levaram a avanços substanciais. Todas as técnicas foram bem sucedidas em seguir os passos propostos, mas nenhuma manteve 100% de precisão em relação às instruções individuais, com todas tendo alguma margem de erro nas mesmas etapas destacadas

anteriormente. Surpreendentemente, o mesmo resultado se manteve mesmo utilizando o modelo GPT-4, entre técnicas simples e avançadas de *prompting*.

Os testes iniciais com a API da OpenAI permitiram verificar a capacidade dos modelos fundacionais e ao mesmo tempo sua inconsistência. Em relação ao requisito 2, permitiram validar que a capacidade dos modelos de se comportar como tutores em relação à questões do ENEM, mas demonstraram que obter um padrão consistente de resposta demandaria esforços adicionais no desenvolvimento. Diante desse resultado, optamos por seguir para a elaboração dos outros aspectos estipulados para depois retornar ao problema da inconsistência.

Para a base de dados a ser utilizada no protótipo, foram acessadas as provas do ENEM mais recentes disponíveis no site do INEP, de 2022 e 2023, extraídos os textos dos enunciados, alternativas e respostas de gabarito. Foram utilizadas apenas as questões que não envolviam interpretação de imagens ou tabelas a fim de simplificar o processo de desenvolvimento, uma vez que lidar com imagens não foi estipulado como requisito do sistema inicial. Os textos foram organizados em um arquivo *csv* com as seguintes estruturas de atributos para serem testadas:

- 1: (i) prova ENEM; (ii) número da questão; (iii) tema da questão; (iv) enunciado da questão; (v) alternativas da questão; (vi) alternativa correta;
- 2: (i) prova ENEM; (ii) número da questão; (iii) tema da questão; (iv) enunciado da questão; (v) a (ix) alternativas da questão uma por coluna; (x) alternativa correta;
- 3: (i) prova ENEM; (ii) número da questão; (iii) tema da questão; (iv) enunciado da questão e alternativas; (v) alternativa correta;

Figura 8 – Base de dados utilizada para validação da hipótese técnica

	A	B	C	D	E	F	G	H	I	J	K
1	ordem	= ano	= tipo	= numero_questao	= enunciado	= alternativa_i	= alternativa_j	= alternativa_k	= alternativa_l	= alternativa_m	= gabarito
2	1	2023	Linguagens	3	3\nThings We Carry on the Sea\nWe carry tears in our eyes: good-bye father, good-b A risco de choque B impacto do en C importância de D existência de E necessidade de F diversidade de G convergência I H característica C tendência estil D discriminação E censura music B						
3	2	2023	Linguagens	1004	4\nSpanlish\npués estoy creando Spanlish\nbicultural systems\nscientific lexicog A convergência I B característica C tendência estil D discriminação E censura music B						
4	4	2023	Linguagens	1003	3\nCaramelos\" em sus suelos\nLas tierras de España, tu vista enamorarán\nnsus gente A característica B beleza das paix C variedade de D relação entre E gastronomia ci E						
5	5	2023	Linguagens	1004	4\nTécnicas de manipulación y el resultado\nManipular es sembrar en la conciencia A amadurece tal B busca mascara C rejeita questo D aproxima-se d E faz-se presente B						
6	6	2023	Linguagens	1005	5\nQue quede claro\nComo es posible que se ciernen tantas bocas, tantos ojos,\nA indignação dia B violência cara C preconceito D descendo da pi E falta de segura A						
7	7	2023	Linguagens	6	6\nA sessão do Comitê Olímpico Internacional (COI)\naprovou uma mudança históri A unificação do I B aproximação C junção do lem D associação ent E vinculação ent D						
8	8	2023	Linguagens	7	7\nMais iluminada que outras\nTenho dois setos, estas duas coisas, duas mãos que n A revelam as ma B questionam o C reproduzem pi D sugerem uma E mimetizam, na A						
9	9	2023	Linguagens	10	10\nNé a interferência de contas falsas em discussões\npolíticas nas redes sociais já A controle da ati B desenvolverme C flexibilização D necessidade de E identificação d E						
10	10	2023	Linguagens	11	11\nMaiô foi colorido de amarelo, e o foi porque\nmundialmente amarelo é a cor cc A enumerar as c B contextualizar C divulgar dados D sensibilizar o p E restringir os pr D						
11	11	2023	Linguagens	12	12\nAinda daquela vez pude constatar a bizarrice dos\nncostumes que constituiam a: A opção por terr B crítica social in C descrição impr D presença de u E ambigüidade de i C						
12	12	2023	Linguagens	13	13\nGrassol da madrugada\nTeu dedo curioso me segue lento no rosto\nos sulcos, A vergonha das i B indecisão em f C serenidade sec D frustração cau E disponibilidade d C						
13	13	2023	Linguagens	14	14\nDâo Lalâo\nDo povoado do Ó, ou dos sítios pert, alguém\nprecisava urgente: A qualidade do s B estabilidade d C ineditismo do D jeito singular c E dificuldade de D						
14	14	2023	Linguagens	15	15\nAs cinzas do Museu Nacional, no Rio de Janeiro,\nconsumido pelas chamas no r A exige a retom B representa dai C impossibilita o D resulta na exti E inviabiliza o es B						
15	15	2023	Linguagens	16	16\nMandioca, macaxeira, aipim e castelinha são nomes\nndiferentes da mesma plar A passa por fén B apresenta vari C sofre mudanç D diferença-se e E é ininteligivel f A						
16	16	2023	Linguagens	17	17\nComo é bom reencontrar os leitores da Revista da\nncultura nor meio de uma n A perfil de nihili B onnisciente dn C exclusividade d D identidade da F nadir editori D						

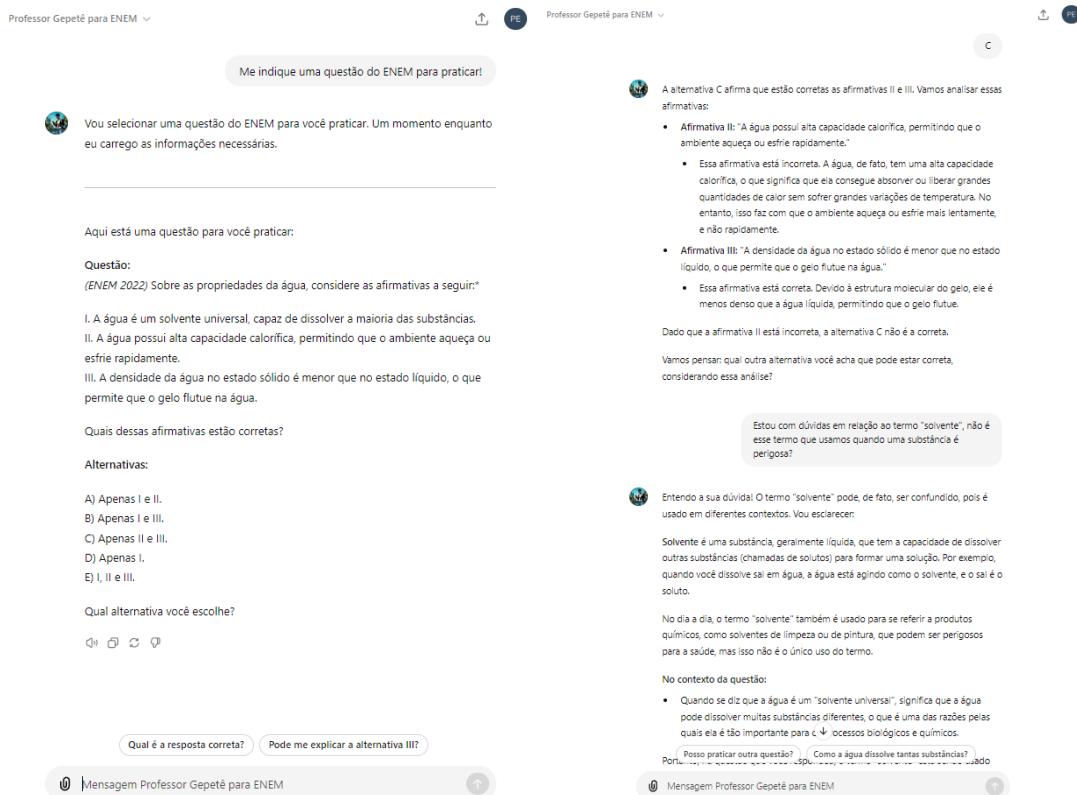
Fonte: elaboração própria.

Em vez de proceder diretamente para o desenvolvimento de um sistema RAG ou de um framework que a incorpora, foi decidido experimentar o produto “GPTs” oferecida pela empresa OpenAI. GPTs são *Chatbots* personalizáveis baseados nos modelos fundacionais da OpenAI, em que o usuário é capaz de especializar o comportamento do modelo através de

*prompting*, do carregamento de arquivos e da configuração de determinação de ações. A intenção aqui simplificar a validação da nossa hipótese, poupano recursos. O arquivo foi carregado e foi estipulado um prompt simples oferecendo instruções de como o modelo deveria se comportar.

O que se verificou foi surpreendente: além da vantagem de fazer referência a um arquivo com questões do ENEM muito mais extenso com precisão, a capacidade do GPT de manter uma consistência entre respostas também foi superior em relação aos testes a partir de requisição de APIs. É provável que esse resultado seja reflexo de um processo de desenvolvimento mais apurado e que permite o modelo seguir instruções com mais fidelidade em relação aos prompts na API.

Figura 9 – Experiência de uso com GPT da OpenAI configurado para ser um tutor para o ENEM



Fonte: elaboração própria.

Uma vez que o objetivo desse projeto é validar a hipótese sobre a viabilidade de utilizar as capacidades de modelos fundacionais por meio de um sistema de prompts efetivos, foi possível validar essa hipótese com os GPTs juntamente a uma base contendo questões do ENEM. O modelo GPT-4 através do produto GPTs foi capaz de incorporar o comportamento de um tutor em relação às questões do ENEM, que avalia as respostas dos usuários, mas não se

limita a oferecer as respostas corretas. O modelo demonstrou ser capaz de fornecer conteúdo útil e direcionar o usuário na direção correta, além de resolver dúvidas específicas, sem ter sido necessário nenhum tipo de treinamento e apenas prompts efetivos em um sistema generalista. Isso oferece uma resposta satisfatória em relação à primeira hipótese, validando a viabilidade técnica da solução proposta no nível do comportamento do modelo.

#### **4.3.2 Hipótese 2 (demanda)**

Para validar a hipótese: “estudantes em preparação ao ENEM sofrem da falta de opções de baixo custo para tirar dúvidas sobre a prova e há interesse em contar com uma ferramenta automatizada baseada em IA que atenda a essa necessidade” é necessário formular hipóteses específicas sobre o público alvo mais adequado, apresentar a solução ou conceito e colher algum tipo de dado que seja capaz de capturar a informações sobre o interesse do público em relação à solução proposta.

A solução visa um público em fase de preparação para o vestibular durante ou após a conclusão do ensino médio. Isso indica uma idade de estudantes de segundo ou terceiro ano do ensino médio, e que pode se estender indefinidamente, já que não há restrição de idade para ingressar no ensino superior. Contudo, dado um foco em atender pessoas mais jovens, adotar uma margem de idade mais estrita ajuda a simplificar experimentos possíveis e a aumentar sua eficácia, considerando os custos envolvidos. Podemos, desse modo, adotar uma janela de idade conjuntamente com escolaridade que vai do segundo ano do ensino médio até 2 anos após a idade média de sua conclusão.

Tendo em vista o problema da desigualdade de oportunidades, é essencial que a solução seja interessante para um público predominantemente de baixa renda. Nesse ponto, é possível usar como referência o salário mínimo (R\$ 1302,00 em valores atuais) e critérios de renda *per capita* por domicílio. Supondo que o usuário potencial não trabalhe (o que pode se mostrar em muitos casos falso), isso configura um cenário em que há menos provedores no domicílio que membros, o que pode resultar até numa fração de um salário mínimo *per capita*. Porém, não faz sentido restringir tanto o público, de modo que pode-se considerar um espectro que vai até o valor mínimo considerado para que uma pessoa seja considerada de classe média no Brasil. Assim, um experimento é programado tendo como alvo uma seleção de público com as seguintes características:

1. Idade entre 16 e 20 anos;

2. Escolaridade ensino médio incompleto ou completo
3. Faixa de renda < 2 salários mínimos per capita no domicílio

Isso pode ser feito de algumas maneiras, mas uma bastante interessante é o uso de mídia paga em redes sociais, como o Instagram. O Instagram é uma rede social amplamente utilizada no Brasil, estando apenas atrás do WhatsApp e YouTube e contando com 113 milhões de usuários (RDStation, 2024). Em relação a estas, o Instagram oferece a possibilidade de propagandas estáticas, o que facilita a elaboração de uma campanha por nós. Além de oferecer uma capilaridade alta, esse sistema é digital e oferece diversas métricas que permitem constatar adesão à proposta. Idealmente, esse processo contaria com experimentação e testes a partir de diferentes chamadas e identidades visuais. Para identificar potencial de interesse, é possível utilizar o MVP via GPT, ou colher alguma sinalização de interesse do público, como e-mail para lista de espera ou respostas a um formulário que explica a solução. A partir de uma comunicação honesta e clara com relação à proposta e da exposição a um público bem delineado, tal experimento poderá render informações valiosas para a continuidade do projeto e mesmo para a obtenção de financiamento.

Outra possibilidade é a divulgação presencial, com contato direto com usuários potenciais. Escolas públicas de nível médio podem ser uma ótima opção nesse sentido, e alguns instrumentos diferentes podem ser utilizados. Um deles é a conversa direta com estudantes do perfil elaborado e da aplicação de um questionário simples desenhado para validar o interesse em relação à proposta. Existem várias metodologias nesse sentido, como por exemplo as apresentadas no livro “*Talking to Humans: Success starts with understanding your customers*”, por Giff Constable e Eric Ries (2014). Dentro da divulgação presencial, pode ser incluída a produção de um flyer com um *QR-code* que leve ao MVP ou a uma lista de interesse.

Estes experimentos podem ser conduzidos com custos baixos, e a realização de pelo menos um deles deve trazer informações para aperfeiçoar o entendimento do público alvo sugerido e do seu potencial e interesse em relação à solução proposta.

#### **4.3.3 Hipótese 3 (efetividade)**

Como discutido anteriormente, essa hipótese é a mais complexa de se avaliada, ao mesmo tempo que é essencial. Por mais que seja viável utilizar LLMs de ponta para criar explicações para o ENEM, e que haja interesse do público em utilizar um sistema automatizado

como ferramenta de preparação, isso pouco importa se a solução não for útil no sentido de melhorar o desempenho de seus usuários no exame. Ela se torna ainda mais relevante quando consideramos a intenção do projeto ser alvo de recursos públicos, que devem ser empregados com muita cautela e responsabilidade.

O primeiro passo para validar a hipótese de entrega de valor é o aprimoramento da solução inicial. Seria inocente considerar que um MVP como o desenhado para validar a hipótese 1 poderia ser suficiente para a finalidade de oferecer uma prática eficaz ao ENEM. Por mais que ele seja capaz de explicar as questões por diferentes pontos de vista e tirar dúvidas, isso é feito de maneira segmentada e não oferece nenhum tipo de direcionamento ao estudante. Um dos componentes essenciais destacados na solução proposta neste trabalho consiste num sistema recomendador que leve em conta o estágio de desenvolvimento do estudante em relação a sua capacidade de responder à prova. Incorporando uma acepção mais clássica de MVP, deve ser produzido um sistema com as características fundamentais propostas para efetivar sua entrega de valor e testado junto a usuários reais. Já é um sistema que exige um processo de desenvolvimento muito mais elaborado, que será especificado na seção seguinte.

Para validar a efetividade da solução proposta, é recomendável seguir o método científico e desenhar um experimento de maneira a isolar o efeito do uso da solução. No plano ideal, obter a adesão de um grupo grande de estudantes dispostos a testar a solução a partir de certas prescrições de assiduidade e a ser observado ao longo de no mínimo o período que compreende a realização de 2 ENEMs; separar grupos de teste e de controle mais similares possíveis através de características ou aleatorização e medir a diferença de desempenho entre os grupos. O mundo dos negócios não depende de um rigor tão grande, mas ainda assim essa prescrição serve de referência sobre o que é realmente necessário para sustentar a validade de uma solução como a proposta.

Diante das limitações de recursos que se impõe, para justificar o seguimento do projeto após a elaboração do MVP, será suficiente contar com um grupo reduzido de estudantes em fase de preparação para o ENEM (ao longo de um ano) com interesse em ser os primeiros a testar uma solução inovadora. Esses *early-adopters* precisarão estar em contato com a equipe responsável e oferecer feedbacks através de questionários estipulados com a finalidade de atender aos requisitos teóricos definidos. A avaliação mais precisa poderia entrar numa etapa posterior, já no âmbito da implementação pública.

#### **4.3.4 Hipótese 4 (viabilidade econômica)**

A hipótese 4 se trata da viabilidade econômica da solução, e esta tem a ver em última análise com os custos e benefícios em etapas posteriores de amadurecimento, que não há como ser avaliados previamente. Em relação à viabilização do projeto dentro de um modelo B2G, porém, vale considerar que ela depende de aproximações com a iniciativa pública e acompanhamento dos editais lançados por instituições públicas que atuam com projetos deste perfil, como as mencionadas na seção 3 e outras.

No anexo A, foi inserido edital lançado pela Fundação Itaú para projetos de Inteligência Artificial aplicada à educação, “IA para Educação”, voltado para tecnologias que sirvam o ensino público básico. Este projeto se adequa perfeitamente a esse edital, atendendo aos critérios definidos nele, e ao mesmo tempo os recursos disponibilizados de até R\$ 200.000,00 seriam suficientes para desenvolver uma solução funcional dentro da especificação que se segue.

A expectativa é que surjam outras oportunidades como essa, em que o projeto possa ser apresentado e selecionado.

### **4.4 Especificação do MVP e orçamento inicial**

#### **4.4.1 Visão geral**

O desenvolvimento de uma plataforma de tutoria automatizada para o ENEM com o uso de Inteligência Artificial Generativa apresenta oportunidades e desafios significativos. Há um enorme potencial em uma solução deste tipo de garantir possibilidades de uma preparação flexível e otimizada para o exame, mas existem muitos componentes decisivos da qual depende sua eficácia. O desenvolvimento de MVP que gere efetivamente este valor depende da integração de um banco de questões do ENEM a uma IA efetiva em sua capacidade de atender a dúvidas e oferecer direcionamentos úteis no contexto de questões particulares e a um sistema capaz de fornecer recomendações de questões adequadas ao progresso dos estudantes com níveis de habilidade diferentes em competências diferentes.

Nesta especificação, serão abordados os principais aspectos do projeto de um MVP que será iterativamente construído através de ciclos de desenvolvimento ágil. A solução utiliza um modelo fundacional, o GPT-4, para gerar explicações detalhadas e respostas personalizadas, enquanto a recomendação de questões é abastecida com dados referentes às provas do ENEM

cedidos pelo INEP. O processo de desenvolvimento será conduzido por uma equipe técnica formada por um desenvolvedor *fullstack*, um engenheiro de dados e um designer de UX, idealmente com experiência em projetos educacionais.

#### **4.4.2 Base de Dados e Dataset "ENEM Challenge"**

A base de dados utilizada no MVP será composta por questões do ENEM extraídas do *dataset "ENEM Challenge"*, que abrange as edições do exame de 2009 a 2017. Esse *dataset* contém as questões, alternativas de resposta e as respectivas respostas corretas, além de algumas outras variáveis. É uma fonte valiosa para ter como ponto de partida uma base consistente e confiável das questões do ENEM, podendo ser complementada posteriormente com extração de dados das provas de 2018 a 2023 através das versões disponibilizadas no site do INEP.

O processamento dessas questões será feito utilizando a linguagem Python, empregando bibliotecas como Pandas para a manipulação de dados. As questões serão armazenadas em um banco de dados relacional (como PostgreSQL ou MySQL) que suportará a indexação e recuperação de questões conforme as regras de recomendação definidas

#### **4.4.3 Sistema Recomendador de Questões**

O sistema recomendador de questões será um componente muito importante da plataforma de tutoria. Ele se baseará em dados cedidos pelos usuários sobre suas aptidões em diferentes competências em um primeiro momento, e depois será abastecido com os dados referentes ao uso dos estudantes da solução. Seu desenvolvimento será enormemente facilitado pela existência de dados públicos sobre os itens de provas do ENEM, com os parâmetros da TRI e de competências e habilidades avaliadas em cada questão.

##### **4.4.3.1 Sistema Recomendador Baseado em TRI**

O sistema recomendador de questões será desenvolvido com base nos parâmetros da Teoria de Resposta ao Item e nas informações detalhadas sobre as competências e habilidades avaliadas em cada questão, conforme fornecidas nos dados cedidos pelo INEP. Essa abordagem permitirá oferecer aos estudantes uma experiência personalizada e adaptativa, alinhada ao seu nível de habilidade em cada competência específica.

Inicialmente, os estudantes fornecerão informações sobre sua autopercepção de aptidão em diferentes competências por meio de um formulário. Embora essa autopercepção possa não ser totalmente precisa, servirá como referência para as primeiras recomendações em cada competência, especialmente na ausência de dados de uso prévio. Após o primeiro uso, as recomendações serão baseadas predominantemente nos dados de desempenho do próprio estudante, reduzindo a importância das informações iniciais.

As recomendações serão oferecidas em blocos de questões, de acordo com a disponibilidade de tempo informada pelo usuário. O sistema calculará o número de questões apropriado para cada sessão, utilizando a média de tempo por questão do ENEM como referência. O nível de dificuldade das questões será utilizado como proxy para estimar o tempo necessário, buscando equilibrar o conjunto de questões para que se adeque ao tempo disponível.

As questões que compõem cada bloco serão selecionadas a partir de uma distribuição de probabilidade na variável de dificuldade, centralizada em um nível de dificuldade orientado pelos usos anteriores do estudante. Essa centralização será ajustada de acordo com o desempenho do estudante em cada competência: acertos indicam um aumento no nível de habilidade estimado, resultando em recomendações de questões mais desafiadoras; erros sugerem uma diminuição, levando a recomendações de questões menos complexas. Dessa forma, as recomendações manterão um equilíbrio entre desafiar o estudante e evitar frustrações, além de refletir a diversidade encontrada nas provas do ENEM.

As distribuições de dificuldade serão calculadas separadamente para cada competência, permitindo que o estudante progride em diferentes ritmos em cada área de conhecimento. Isso possibilita um acompanhamento mais preciso do desenvolvimento individual e uma personalização mais efetiva das recomendações.

Ao final de cada bloco de questões, o sistema atualizará as estimativas de habilidade do estudante em cada competência, com base em seu desempenho. Essa atualização não ocorrerá entre questões individuais, mas entre práticas que consistem em blocos de questões, garantindo uma quantidade de dados suficiente para ajustes significativos. Com isso, as recomendações futuras serão refinadas, tornando-se cada vez mais alinhadas ao perfil de aprendizado do estudante.

Esse direcionamento inicial do sistema recomendador poderá ser aperfeiçoado à medida que testes com usuários forem realizados e com o apoio de pedagogos, visando aumentar a efetividade da solução. A colaboração com especialistas em educação permitirá ajustar os critérios de recomendação e incorporar estratégias pedagógicas que potencializem o aprendizado.

#### 4.4.4 Integração com GPT-4 para Geração de Respostas

O GPT-4 será o modelo fundacional responsável por gerar respostas e explicações detalhadas para as questões recomendadas. O modelo será integrado à plataforma através da API da OpenAI, e seu uso será acionado para mediar todas as interações com o usuário, seja para dar boas-vindas, apresentar questões selecionadas pelo sistema recomendador, fornecer feedbacks com relação a respostas ou oferecer uma explicação detalhada para a questão respondida.

Essa integração será feita utilizando a linguagem Python, conectada diretamente ao *back-end* da aplicação. O modelo GPT-4 receberá como input a questão selecionada e o contexto da solicitação do usuário, gerando uma resposta clara e personalizada. A escolha do GPT-4 foi baseada em sua capacidade de gerar explicações coerentes e adaptativas, essenciais para uma experiência educacional de alta qualidade.

Para uma integração eficaz do GPT-4 na solução, será necessário um processo de desenvolvimento por *prompting* baseado em experimentações e iterações, tendo em vista critérios estabelecidos para o comportamento do modelo. As validações anteriores nos deixam confiantes no sucesso dessa etapa ao mesmo tempo que previnem com relação a sua imprevisibilidade, sendo mais difícil estimar a quantidade de iterações necessária para a obtenção de um resultado satisfatório.

#### 4.4.5 Interface de Usuário e Personalização

A interface será desenvolvida em React e terá como base princípios de design focados em usabilidade educacional. Utilizaremos um formato de *Chatbot* ao qual usuários estão acostumados por conta dos populares aplicativos de mensagens. O objetivo é proporcionar uma experiência simples e intuitiva para o usuário, sem sacrificar a sofisticação necessária para personalização e recomendação precisa de questões, que ocorrerá como que em um diálogo.

##### 4.4.5.1 Formulários de Personalização

No caso do sistema recomendador alternativo, a interface incluirá formulários de entrada para coletar informações sobre o perfil do estudante. Esses dados serão armazenados no banco de dados e utilizados pelo *back-end* para personalizar as recomendações. Será

necessário decidir os dados prioritários a serem coletados, e o designer de UX será responsável por criar uma interface que seja fácil de usar e que incentive a interação dos estudantes, sem sobrecarregá-los com perguntas excessivas.

#### **4.4.6 Desenvolvimento Ágil e iterativo**

O desenvolvimento do MVP se dará a partir de princípios da metodologia Ágil, com sprints de duas semanas para permitir entregas incrementais e testes com usuários reais ao final do processo. Durante o processo de desenvolvimento, as funcionalidades principais serão implementadas e testadas internamente ao final de cada sprint, com ajustes rápidos baseados nos resultados obtidos.

##### **4.4.6.1 Estrutura do Ciclo de Sprints**

O ciclo de desenvolvimento terá quatro sprints ao longo de dois meses:

- **Sprint 1:** Configuração inicial do *back-end* e integração com a nuvem, configuração do banco de dados com integração do *dataset* "ENEM Challenge" e da base do INEP (parâmetros da TRI e competências de cada item), desenho da plataforma através de *wireframes* e protótipos iniciais, início da calibração automática do sistema recomendador. **Testes ao final do sprint:** validação interna da arquitetura, revisão dos *wireframes* e testes iniciais de integração dos dados.
- **Sprint 2:** Desenvolvimento do *frontend*, implementação inicial do sistema recomendador, integração à API do GPT-4 para geração de respostas, criação dos formulários de personalização, elaboração dos critérios de comportamento do modelo com experimentos iniciais de prompting, colaboração com pedagogos para revisão das estratégias educacionais. **Testes ao final do sprint:** testes internos de usabilidade e funcionalidade, ajustes com base no feedback e validação pedagógica.
- **Sprint 3:** Integração completa da solução (*frontend*, *backend* e APIs), desenvolvimento da atualização dos dados dos usuários para as recomendações, aperfeiçoamento do comportamento do modelo através de engenharia de prompt visando a experiência do usuário, implementação de mecanismos de análise (ferramentas de monitoramento de uso), planejamento e execução de testes de desempenho e segurança. **Testes ao final do sprint:** testes com usuários reais, coleta de feedback para melhorias, testes de carga e segurança, ajustes conforme necessário.

- **Sprint 4:** Ajustes finais e otimizações no sistema, finalização da calibração automática do sistema recomendador, preparação para o lançamento (*deploy*, configuração de servidores, escalabilidade), desenvolvimento de materiais de apoio (tutoriais, FAQs, onboarding), elaboração do plano de lançamento, incluindo estratégias de marketing e suporte ao usuário. **Testes ao final do sprint:** testes finais de aceitação, validação completa do sistema e preparação para o lançamento.

#### **4.4.7 Infraestrutura Tecnológica**

A infraestrutura tecnológica será baseada em serviços de hospedagem em nuvem, como AWS ou Google Cloud, para garantir escalabilidade e flexibilidade. O uso da API do GPT-4 será dimensionado conforme a demanda dos usuários, com uma provisão inicial de R\$ 3.000/mês para o uso da API.

Além disso, ferramentas de monitoramento e desenvolvimento, como Sentry para acompanhamento de erros e GitHub para controle de versão, serão utilizadas para garantir a integridade do processo de desenvolvimento.

#### **4.4.8 Orçamento e Custos Estimados**

Tendo em vista essa especificação, estimamos um custo para desenvolvimento do MVP e teste junto a usuários de **R\$ 133.000**. Esse valor é uma aproximação que pode sofrer alterações dependendo da formação da equipe e de desafios não previstos, mas nos permite orientar a busca por editais ou investimentos na faixa de R\$ 120.000 – R\$ 150.000 que prevejam apenas o desenvolvimento inicial da solução.

##### **4.4.8.1 Custo de Desenvolvimento**

A equipe de desenvolvimento será composta por três profissionais (designer de UX, desenvolvedor *fullstack*, engenheiro de dados), cada um com um salário mensal aproximado de R\$ 10.000. O custo estimado com recursos humanos em desenvolvimento ao longo de três meses de trabalho será de R\$ 90.000.

##### **4.4.8.2 Infraestrutura Tecnológica e Ferramentas Pagas**

- Servidores e hospedagem em nuvem: R\$ 2.000/mês (previstos 3 meses adicionais).
- Uso da API GPT-4: R\$ 3.000/mês (previstos 3 meses adicionais).

- Ferramentas pagas e serviços adicionais: R\$ 500/mês (previstos 3 meses adicionais). Custo estimado com infraestrutura tecnológica: R\$ 33.000.

#### **4.4.8.3 Budget de Divulgação**

Um orçamento de R\$ 10.000 será destinado para a divulgação da plataforma ao final do período de desenvolvimento. Esse valor será utilizado para campanhas digitais em redes sociais e outras plataformas de divulgação.

#### **4.4.9 Indicação de um *roadmap***

Por fim, para pensar nos direcionamentos futuros desta proposta de sistema de tutoria para o ENEM, vale retomar as colocações de Skinner com os requisitos para uma “máquina de ensino” que simule um “bom tutor”, que também nos aproxima do problema Bloom-2-Sigma. A prescrição de Skinner pode ser tomada como base para identificação de algumas funcionalidade ideais da solução:

- 1) Ser mais engajante e sustentar a atenção do usuário;
- 2) Explorar diferentes questões da base referentes a um mesmo assunto, e apenas mudar quando o estudante demonstrar ter dominado este assunto;
- 3) Oferecer questões no nível de dificuldade adequado para a progressão do estudante;
- 4) Adaptar as técnicas de explicação para a melhor compreensão do aluno individual;
- 5) Oferecer algum estímulo com impacto positivo sobre o estudante em caso de acerto ou demonstração de aprendizado;

O *roadmap* da solução seguirá essa orientação, com destaque para o aprimoramento do engajamento e a adoção do modelo de *mastery learning*.

Para tornar a plataforma mais envolvente e sustentar a atenção dos estudantes, uma possibilidade interessante é implementar técnicas de gamificação que recompensem o progresso e incentivem a continuidade no estudo. Recursos como a contagem de dias consecutivos de uso (*streak*) e a criação de desafios personalizados podem ser desenvolvidos para motivar os usuários. Além disso, recompensas visuais e estatísticas de desempenho atreladas aos acertos nas questões contribuirão para manter os estudantes engajados.

Já no contexto de *mastery learning*, a plataforma deve garantir que o estudante explore diferentes questões sobre o mesmo assunto até alcançar a maestria. Para isso, a inclusão de *Retrieval-Augmented Generation* (RAG) torna-se relevante, pois permitirá organizar as

questões por tópicos além da por nível de dificuldade. Dessa forma, a plataforma poderá oferecer uma progressão natural das questões mais simples para as mais complexas dentro de um mesmo assunto, ao mesmo tempo adaptando-se à curva de aprendizado de cada aluno e trazendo condições para maior consistência. O modelo poderá também gerar questões adicionais ou práticas personalizadas, com o devido cuidado para evitar alucinações, a fim de garantir a precisão e a qualidade do material oferecido.

## CONCLUSÃO

A motivação inicial deste trabalho foi a aplicação prática dos estudos teóricos sobre *Large Language Models* (LLMs), uma tecnologia que o autor considera revolucionária, no contexto do MBA em Inteligência Artificial oferecido pelo ICMC-USP. O objetivo foi desenvolver uma aplicação que aproveitasse as principais características dos LLMs, aliando sua versatilidade a um impacto positivo e significativo na vida de pessoas.

Dessa proposta surgiu a ideia de utilizar a versatilidade própria aos LLMs para tirar dúvidas sobre o ENEM, aproveitando suas capacidades para o contexto educacional. Embora a concepção do autor valorize o aspecto relacional da educação, a afetividade do contato humano e o papel fundamental do professor, reconhece-se que a Inteligência Artificial Generativa, assim como qualquer tecnologia inorgânica, é impotente em relação a estes aspectos. No entanto, essa visão contrasta com a lógica dos exames padronizados, que se baseiam em critérios objetivos e padronizados, contexto no qual sistemas de IA, como os LLMs, mostram-se extremamente adequados.

Exames padronizados, por sua própria natureza, demandam precisão e repetição de padrões — características em que a IA e os LLMs se destacam. Dada a importância da prática na preparação para esses exames, a possibilidade de oferecer prática assistida a baixo custo através dos LLMs representa uma aplicação promissora. Quando integrada aos avanços empíricos de sistemas de ensino adaptativos, essa tecnologia pode se mostrar de amplo impacto. Além disso, com a evolução constante dos modelos, as limitações atuais tendem a ser superadas com o tempo.

Os primeiros passos no sentido do desenvolvimento prático da solução confirmaram a pertinência da tecnologia para a tarefa proposta. No entanto, essa validação inicial suscitou questões sobre "para quem" e "para que" serviria essa aplicação, levando à constatação de que o público-alvo tratado aqui — estudantes de baixa renda — são os que mais precisam dessa solução, mas não podem pagá-la. Em contrapartida, aqueles que poderiam arcar com o custo dessa tecnologia já dispõem de outros recursos. Esse paradoxo orientou a mudança de enfoque do projeto, que passou a buscar sustentar a proposta de uma solução inovadora por meio do setor público, deslocando o desafio técnico para um desafio de implementação em um contexto incomum para este tipo de projeto.

Considera-se que o projeto foi relativamente bem-sucedido em atingir seus objetivos. Foi possível definir uma solução de tutoria para o ENEM baseada em IA, demonstrar sua pertinência no contexto social e educacional brasileiro e elaborar um modelo viável para sua

implementação. O trabalho fundamentou-se em pesquisa teórica, que permitiu especificar tecnicamente os pré-requisitos para a entrega de valor da solução, e em uma pesquisa prática, que destacou as principais oportunidades e riscos. Esses elementos foram integrados no desenho de um projeto com maior potencial de sucesso.

O uso do termo "relativamente" reflete a percepção de que ainda há muito espaço para aprimoramentos. Certamente, existem dados mais abrangentes sobre o contexto educacional brasileiro que poderiam embasar melhor a proposta e complementar a execução prática. A pesquisa teórica poderia ser refinada com elementos técnicos adicionais, e a pesquisa prática se beneficiaria de uma análise mais aprofundada de soluções similares, como a Riiid, para extrair aprendizados específicos. Além disso, o avanço nos experimentos propostos, como o de demanda, poderia reduzir ainda mais a percepção de risco associada ao projeto.

Este trabalho representa apenas o início de um projeto com o propósito de promover uma sociedade mais justa e democratizar oportunidades educacionais. Tal objetivo não pode ser atingido por iniciativas individuais, mas apenas por esforços coletivos. Em um mundo orientado pelo desenvolvimento econômico e tecnológico frenético voltado ao interesse privado, o maior mérito deste trabalho talvez resida em sua proposta de um caminho alternativo para enfrentar desafios coletivos. Para que esse propósito se concretize e gere impacto real, será essencial construir parcerias sólidas e abrangentes. Apenas o futuro dirá.

## REFERÊNCIAS

ALMEIDA, Thales Sales et al. Sabiá-2: A New Generation of Portuguese Large Language Models. 2024. Disponível em: <https://arxiv.org/abs/2403.09887>. Acesso em: 02 abr. 2024.

BLOOM, Benjamin S. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher*, v. 13, n. 6, p. 4-16, 1984.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. 3,9 milhões estão inscritos no Enem 2023. Disponível em: <https://www.gov.br/inep/pt-br/assuntos/noticias/enem/3-9-milhoes-estao-inscritos-no-enem-2023>. Acesso em: 3 jul. 2024.

BRAZIL JOURNAL. UOL EdTech compra Passei Direto. Disponível em: <https://braziljournal.com/uol-edtech-compra-passei-direto/>. Acesso em: 3 ago. 2024.

CONSTABLE, Giff; RIES, Eric. Talking to humans: sucesso na entrevista com clientes para startups, projetos e inovações. [S.l.]: Giff Constable, 2014.

CHUNG, Juliet. A South Korean Company Uses AI to Help Students Ace Their Exams. *The New York Times*, 27 dez. 2022. Disponível em: <https://www.nytimes.com/2022/12/27/business/ai-education-app-riiid.html>. Acesso em: 13 ago. 2024.

DATA REPORTAL. Digital 2023: Global Overview Report. Disponível em: <https://datareportal.com/reports/digital-2023-global-overview-report>. Acesso em: 2 abr. 2024.

EMBRETSON, S. E.; REISE, S. P. Item Response Theory for Psychologists. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

ENDEAVOR BRASIL. Como o Descomplica fez do aprendizado contínuo uma estratégia de crescimento. Disponível em: <https://endeavor.org.br/historia-de-empreendedores/como-o-descomplica-fez-do-aprendizado-continuo-uma-estrategia-de-crescimento/>. Acesso em: ago set.

FERRAZ, Tainá. Da Bahia para o futuro: edtech quer ensinar programação para todos. *Fast Company Brasil*, 14 ago. 2023. Disponível em: <https://fastcompanybrasil.com/impacto/da-bahia-para-o-futuro-edtech-quere-ensinar-programacao-para-todos/>. Acesso em: 13 ago. 2024.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. Relatório de resultados do Saeb 2019: volume 1: 5º e 9º anos do Ensino Fundamental e séries finais do Ensino Médio [recurso eletrônico]. Brasília, DF: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2021. 245 p.: il.

JI, Z.; LEE, N.; ISLAM, F.; MADOTTO, A.; MOU, L. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys (CSUR)*, 55(12), 2023.

KHAN ACADEMY. What is the history of Khan Academy. *Khan Academy Support*, 2023. Disponível em: <https://support.khanacademy.org/hc/en-us/articles/202483180-What-is-the-history-of-Khan-Academy>. Acesso em: 13 ago. 2024.

KHAN, Salman. Brave new words: how AI will revolutionize education (and why that's a good thing). New York: Viking, 2024.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. *Nature*, v. 521, n. 7553, p. 436-444, 2015.

LEWIS, Patrick et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020). 2020.

LYONS, Kim. Google discloses its acquisition of mobile learning app Socratic as it relaunches on iOS. *TechCrunch*, 16 ago. 2019. Disponível em: <https://techcrunch.com/2019/08/16/google-discloses-its-acquisition-of-mobile-learning-app-socratic-as-it-relaunches-on-ios/>. Acesso em: 13 ago. 2024.

LORD, F. M. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum Associates, 1980.

McKINSEY & COMPANY. The state of AI in 2023: Generative AI's breakout year. Disponível em: <https://www.mckinsey.com>. Acesso em: 2 abr. 2024.

MUGNAINI JUNIOR, A. N.; CUNHA, M. S. S. IMPACTO DAS COTAS NO DESEMPENHO DE ESTUDANTES NO CURTO E LONGO PRAZOS. Planejamento e Políticas Públicas, [S. l.], n. 64, 2023. DOI: 10.38116/ppp64art2. Disponível em: //www.ipea.gov.br/ppp/index.php/PPP/article/view/1414. Acesso em: 14 ago. 2024.

NUNES, Desnes et al. Evaluating GPT-3.5 and GPT-4 Models on Brazilian University Admission Exams. 2023. Disponível em: <https://arxiv.org/abs/2303.17003>. Acesso em: 02 abr. 2024.

OECD (2018), A Broken Social Elevator? How to Promote Social Mobility, OECD Publishing, Paris, <https://doi.org/10.1787/9789264301085-en>.

OECD (2023), PISA 2022 Results (Volume I): The State of Learning and Equity in Education, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/53f23881-en>.

PIRES, Ramon et al. Evaluating GPT-4's Vision Capabilities on Brazilian University Admission Exams. 2023. Disponível em: <https://arxiv.org/abs/2311.14169>. Acesso em: 02 abr. 2024.

RIES, Eric. A startup enxuta: como os empreendedores atuais utilizam a inovação contínua para criar empresas extremamente bem-sucedidas. Rio de Janeiro: Elsevier, 2012.

THE GUARDIAN. ChatGPT: 100 million users make OpenAI's invention fastest-growing app. 03/02/2023. Disponível em:  
<https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app>. Acesso em: 2 abr. 2024.

TOLEDO, Bruna. Cotas se equiparam às vagas amplas pela 1ª vez nas universidades estaduais. Folha de S.Paulo, São Paulo, 04 mar. 2024. Disponível em:  
<https://www1.folha.uol.com.br/educacao/2024/03/cotas-se-equiparam-as-vagas-amplas-pela-1a-vez-nas-universidades-estaduais.shtml>. Acesso em: 13 ago. 2024.

VAN DER LINDEN, W. J.; HAMBLETON, R. K. (Eds.). Handbook of Modern Item Response Theory. New York: Springer, 1997.

SKINNER, B. F. Teaching machines. *Science*, v. 128, n. 3330, p. 969-977, 1958.

SILVEIRA, Igor Cataneo; MAUÁ, Denis Deratani. *University Entrance Exam as a Guiding Test for Artificial Intelligence*. In: Proceedings of the 6th Brazilian Conference on Intelligent Systems (BRACIS), 2017. p. 426–431.

STANFORD UNIVERSITY. AI Index Report 2023 – Artificial Intelligence Index. Disponível em: <https://aiindex.stanford.edu>. Acesso em: 2 abr. 2024.

WEISS, D. J. (Ed.). New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing. New York: Academic Press, 1983.

ZHAO, Wayne Xin et al. A Survey of Large Language Models. 2023. Disponível em: <https://github.com/RUCAIBox/LLMSurvey>. Acesso em: 15 mar. 2024.

## ANEXO A – Edital Itaú IA para a Educação (2024)



### **1. Quem somos**

A Fundação Itaú busca inspirar e criar condições para promover o desenvolvimento de cada brasileiro como cidadão capaz de transformar a sociedade. Por meio das três frentes – Itaú Cultural, Itaú Social e Itaú Educação e Trabalho, a Fundação Itaú dedica programas, ações e articulação com diferentes setores da sociedade para atender às urgências do Brasil contemporâneo. Com uma atuação coerente e sólida, nosso horizonte é fazer sempre e fazer para a equidade, em um exercício vivo e dinâmico que gere legados potentes. No que diz respeito à Inteligência Artificial (IA), um campo da ciência da computação que busca criar sistemas capazes de realizar tarefas que se assemelham à inteligência humana e considerando a perspectiva de aumento da utilização dessa tecnologia, buscamos fomentar projetos de IA para gerar impacto positivo na sociedade por meio da inclusão educacional. Para saber mais: <http://fundacaoitaubr.org.br>

### **2. Edital IA para Educação**

O Edital IA para Educação (“Edital”) é uma iniciativa da Fundação Itaú e tem como objetivo expandir o conhecimento e realizar o acompanhamento da evolução da IA, em especial a IA Generativa, direcionada para a educação básica pública no Brasil. Essa iniciativa se compromete essencialmente com a mitigação das disparidades sociais, a valorização do papel do professor e o engajamento dos estudantes no processo de ensino-aprendizagem. Além disso, visa garantir uma aprendizagem contínua e uma trajetória escolar de excelência, consolidando, assim, a premissa de uma educação integral. Por meio dessa iniciativa, almejamos identificar experiências e iniciativas da produção em IA na educação básica pública no contexto brasileiro, mapear a rede de pesquisadores/cientistas e centros de pesquisas atuantes nesse campo e, por fim, fomentar o desenvolvimento de projetos aplicáveis na educação. A Fundação Itaú busca, por este Edital, fomentar projetos em IA aplicados à educação pública, com foco em pesquisas aplicadas e relacionados a essa temática. Serão aceitos projetos que se enquadrem e preencham os requisitos mínimos da modalidade selecionada na inscrição, sendo elas: (I) Pesquisa Aplicada, (II) Experimentação ou (III) Desenvolvimento de Sistema. Os projetos em questão podem estar em fase de desenvolvimento, desde que tenham sido iniciados a partir de 2022. I - Pesquisa Aplicada: categoria que abrange a realização de pesquisa, inédita ou em andamento, sobre o uso de IA no contexto da educação básica pública brasileira, bem como seus subconjuntos, inclusive IA Generativa. II - Projeto Experimental: categoria que abrange a realização de projetos experimentais relacionados à IA Artificial, incluindo seus subconjuntos, inclusive IA Generativa, com foco no desenvolvimento de estratégias e metodologias para serem testadas no contexto da educação básica pública no Brasil. III - Desenvolvimento de Sistema: categoria que abrange o desenvolvimento de sistemas, aplicativos, APIs ou similares utilizando IA, inclusive seus subconjuntos e IA Generativa, voltados para o contexto da educação básica pública brasileira. Quantos projetos serão contemplados? Não há número mínimo ou máximo definido. As escolhas são de exclusiva atribuição da comissão de seleção da Fundação Itaú, observando-se o limite orçamentário do Edital e os critérios de avaliação dos projetos. **Qual é o valor máximo por projeto contemplado? O valor máximo bruto para projetos inscritos, tanto por pessoas físicas quanto por pessoas jurídicas, é de: a) Pesquisa Aplicada - Até R\$ 50 mil; b) Pesquisa Experimental - Até 100 mil; c) Desenvolvimento de Sistema - Até R\$ 200 mil.** Atenção: o projeto poderá contar com apoio externo, desde que o apoiador não possua critério de exclusividade.

### **3. Quem pode participar**

Poderão se inscrever todas as pessoas físicas ou jurídicas interessadas (“Participantes”) em ações no âmbito da educação e tecnologia, que desejem aprofundar estratégias e projetos de IA, atendendo aos seguintes

critérios: a) Pessoa física (com idade mínima de 18 anos completados até o último dia de inscrição e com CPF ativo); b) Pessoa jurídica e entidades de natureza privada com ou sem fins lucrativos, com sede no Brasil e documentos sociais (contrato social, alterações, estatutos e atas) atualizados e registrados, aptas a firmar contratos e a emitir documentos fiscais. Exemplos: empresas LTDA., Eireli, EPP, MEI, EI, organizações sociais (OS), organizações da sociedade civil de interesse público (OSCIPs), Organizações não Governamentais (ONGs), associações de amigos, associações civis, fundações privadas e similares; c) Ter conta bancária ativa atrelada a CPF ou CNPJ inscritos, com movimentação bancária há, no mínimo, 30 (trinta) dias corridos e que não sejam de incentivos fiscais. Se o Participante for contemplado neste Edital e, no momento da aprovação, sua conta bancária não atender a essas especificações, poderá ser desclassificado a exclusivo critério da Fundação Itaú. Atenção! Não serão aceitas inscrições de: 1. Governos, prefeituras, ministérios, secretarias, órgãos do judiciário e coordenadorias; 2. Membros das casas legislativas (senadores, deputados, vereadores), titulares, vices, adjuntos, assessores diretos e coordenadores, comissionados ou não; 3. Pessoas que possuam impedimento de ordem legal para recebimento de valores por outras fontes, como professores e outros profissionais com regime de dedicação exclusiva ou aposentados por invalidez; 4. Funcionários e estagiários da Fundação Itaú e seus pais ou responsáveis legais, assim como irmãos, filhos, cônjuges e companheiros; 5. Ex-funcionários e ex-estagiários com menos de um ano de desligamento da Fundação Itaú a contar da data de início das inscrições; 6. Funcionários e estagiários do Itaú Unibanco; 7. Pessoas físicas ou organizações que possuam, em seu quadro de colaboradores, funcionários, dirigentes e quaisquer pessoas diretamente envolvidas no planejamento, na avaliação e na execução dessa ação, estendendo-se este requisito inclusive aos parceiros do projeto; 8. Projetos que não envolvam temáticas relacionadas à educação básica; 9. Projetos que não apresentem as autorizações necessárias da Lei de Direitos Autorais, em caso de uso de obras de terceiros; 10. Mais de um projeto por CPF ou CNPJ. A inscrição de pessoas impedidas estará sujeita a desclassificação em qualquer etapa do Edital. Além dos critérios dispostos acima, é possível que, ao longo do processo, outras razões de desclassificação se façam presentes.

## **5. Como funciona a seleção**

Durante o processo de seleção, os projetos inscritos passarão por etapas eliminatórias e serão avaliados de acordo com os critérios mencionados neste Edital e explicitados abaixo, com a participação de especialistas e/ou profissionais em educação e tecnologia. I. TRIAGEM Os Projetos são elegíveis e estão de acordo com os critérios exigidos neste Edital para efetivação da participação? Será realizada a validação dos projetos que cumprirem os requisitos mínimos de elegibilidade e estejam de acordo com as especificações e condições do Edital. Os projetos que passarem pela etapa de validação serão analisados pelo comitê de especialistas designado. II. PARECER TÉCNICO Os projetos atendem aos critérios de seleção? (comitê de avaliação) Os projetos serão submetidos à avaliação por especialistas da área, os quais elaborarão pareceres específicos para cada projeto participante dessa etapa. Essa análise levará em consideração a relevância científica e social da pesquisa/projeto, a amplitude do potencial dos resultados no que diz respeito à melhoria da qualidade da aprendizagem, ao desenvolvimento integral, à equidade e outros efeitos na comunidade escolar, no território e na sociedade. Além disso, será avaliada a sua viabilidade e sua adequação às modalidades previstas no Edital. III. ANÁLISE DE DIREITOS AUTORAIS E HOMOLOGAÇÃO Os documentos enviados estão em conformidade? Será realizada uma avaliação jurídica dos documentos enviados para cada projeto inscrito, incluindo a análise dos direitos autorais relativos aos dados empregados no projeto, bem como do armazenamento, da gestão e do processamento desses dados, sempre em conformidade com a LEI Geral de Proteção de Dados Pessoais (LGPD). Caso necessário, a fundação poderá solicitar documentos e informações complementares, além do cadastramento dos Participantes na plataforma de fornecedores da Fundação Itaú. IV. FORMALIZAÇÃO E CONTEMPLAÇÃO Os projetos atendem a todos os critérios para a formalização do contrato? Os projetos finalistas seguirão para a negociação conduzida pelas áreas de compras e jurídica da Fundação Itaú. Se houver empate, o parecer técnico dos especialistas será utilizado como critério de desempate. A Fundação Itaú reserva-se o direito de não dar devolutivas personalizadas para cada projeto participante sobre as decisões referentes às inscrições, análises e avanço nas etapas ou, ainda, de não fornecer qualquer documento relacionado à análise e ao critério de seleção.

## **6. Critérios de seleção**

Serão aceitos projetos que se enquadrem em uma das três seguintes modalidades, por meio de critérios de qualidade específicos a cada uma delas, quais sejam:

#### Pesquisa Aplicada

- Currículo do proponente e equipe ▪ Título, palavras-chave e resumo ▪ Justificativa da relevância do projeto ▪ Objetivo geral e objetivos específicos do projeto ▪ Qual é a metodologia utilizada no projeto? ▪ Orçamento/Planilha de composição de custos ▪ Referências bibliográficas ▪ Cronograma de execução da pesquisa
- Resultados esperados e produtos a serem desenvolvidos ▪ De que forma esse projeto contribui para a equidade e a redução das desigualdades de raça/cor, gênero e deficiência na educação pública básica brasileira? ▪ Qual é a relação que o projeto estabelece com o uso de IA? ▪ Quais são os públicos atingidos (diretos/indiretos)? ▪ Outras informações julgadas pertinentes para o entendimento da pesquisa

#### Projeto Experimental

- Currículo do proponente e equipe ▪ Título e resumo do projeto ▪ Justificativa da relevância do projeto ▪ Objetivo geral e objetivos específicos do projeto ▪ Qual é a metodologia utilizada no projeto? ▪ Orçamento/Planilha de composição de custos ▪ Cronograma de execução do projeto ▪ Teorias, conceitos e/ou pesquisas que sustentam e justificam a proposta do projeto ▪ Resultados esperados e produtos a serem desenvolvidos ▪ De que forma esse projeto contribui para a equidade e a redução das desigualdades de raça/cor, gênero e deficiência na educação pública básica brasileira? ▪ Qual é a relação que o projeto estabelece com o uso de IA? ▪ Quais são os públicos atingidos (diretos/indiretos)? ▪ Outras informações julgadas pertinentes para o entendimento da pesquisa

#### **Desenvolvimento de Sistema**

- **Currículo do proponente e equipe ▪ Título e resumo do projeto ▪ Justificativa da relevância do projeto ▪ Objetivo geral e objetivos específicos do projeto ▪ Qual é a metodologia de desenvolvimento utilizada no projeto? ▪ Orçamento/Planilha de composição de custos ▪ Cronograma de execução do projeto ▪ Teorias, conceitos e/ou pesquisas que sustentam e justificam a proposta do projeto ▪ Resultados esperados, aplicabilidade e produtos a serem desenvolvidos ▪ Tecnologias e linguagens a serem utilizadas no projeto ▪ De que forma esse projeto contribui para a equidade e a redução das desigualdades de raça/cor, gênero e deficiência na educação pública básica brasileira? ▪ Qual é a relação que o projeto estabelece com o uso de IA? ▪ Quais são os públicos atingidos (diretos/indiretos)? ▪ Outras informações julgadas pertinentes para o entendimento da pesquisa**

Não serão aceitos projetos nos formatos de: – Cursos, oficinas, encontros, debates, palestras, seminários, intercâmbios ou residências com o objetivo de estudo, aperfeiçoamento ou formação; – Mostras, exposições, festivais, feiras ou outros que não se enquadrem nas modalidades e requisitos relacionados a este Edital.

## 7. Sobre os projetos selecionados

Os projetos serão selecionados por uma comissão composta por participantes internos e externos à Fundação Itaú, especialistas na área de Educação e Transformação Digital. A comissão poderá priorizar projetos que versem sobre a IA Generativa. Nas hipóteses de impedimento, doença, força maior ou outras situações que impeçam a participação de algum membro da comissão de seleção, a Fundação Itaú resguarda-se o direito de substituição, com o objetivo de assegurar a equidade nas análises. Meu projeto poderá sofrer alterações durante o processo de seleção? Sim. A comissão de seleção poderá propor alterações no projeto e no orçamento, incluindo a possibilidade de alteração do valor final, com o objetivo de potencializá-lo ou otimizá-lo. O Participante será consultado sobre as sugestões da comissão de seleção – por exemplo: apoio a somente uma etapa do projeto ou aporte de valor diferente daquele solicitado na inscrição. Caso as partes não cheguem a um consenso sobre as alterações propostas, a Fundação Itaú resguarda-se o direito de não incluir o projeto na relação final de contemplados.

## **8. Sobre a divulgação do resultado final**

A divulgação do resultado será no dia 17/10/2024 e será comunicada aos contemplados por telefone e/ou e-mail. A relação de projetos selecionados também será divulgada nos sites institucionais e/ou redes sociais da Fundação Itaú. A Fundação Itaú reserva-se o direito de adiar a divulgação dos resultados – unicamente em situação de força maior. Os projetos selecionados serão contatados pelas equipes de compras e jurídica da Fundação Itaú para prosseguimento do processo de negociação e formalização do contrato. No momento oportuno, a Fundação Itaú informará sobre os procedimentos e critérios mínimos necessários.

## **9. Acompanhamento dos projetos**

Caso o projeto esteja em andamento, a Fundação Itaú poderá realizar o acompanhamento em conjunto com outros agentes envolvidos no contexto da educação básica pública. Esse acompanhamento abordará questões pedagógicas, metodológicas e didáticas, podendo incluir pelo menos uma visita de campo para observação in loco do progresso do projeto.

## **10. Condições de contratação, pagamento e contrapartidas**

**Contratação** O Participante fica ciente de que os contratos serão formalizados com os dados constantes na inscrição e, por essa razão, é importante que os documentos das pessoas físicas e jurídicas sejam válidos, atualizados e autorizados a assumir compromissos jurídicos, bem como a receber valores e emitir documentos fiscais. Não haverá alteração de Participante caso o projeto venha a ser contemplado. Os projetos selecionados deverão ser necessariamente concluídos em até 24 meses a partir da data de assinatura do contrato. Os contratos irão estabelecer, entre outros itens, as responsabilidades das partes, as etapas e o cronograma de realização e acompanhamento do projeto, os requisitos de prestação de contas e a forma e as condições de liberação dos recursos. O cronograma de execução e suas etapas serão o resultado de uma construção entre a proposta de projeto inscrita e as considerações da comissão de seleção e da equipe da Fundação Itaú, sempre em acordo com o Participante. É de exclusiva responsabilidade do Participante contemplado providenciar toda e qualquer documentação necessária para viabilizar o projeto conforme proposto, como alvarás, seguros e, especialmente, as autorizações de direitos de imagem e licenças ou cessões de direitos autorais patrimoniais exigidas para execução e utilização do projeto na extensão e nas modalidades acordadas com a Fundação Itaú. É de inteira responsabilidade do Participante contemplado assegurar os direitos autorais relativos aos dados empregados no projeto, assim como o armazenamento, a gestão e o processamento dos dados durante todas as fases do presente edital, sempre em conformidade com a LGPD. Após a assinatura do contrato, quaisquer alterações nas condições pactuadas estarão sujeitas à aprovação da Fundação Itaú, que também avaliará a necessidade de formalização de termos aditivos.

**Pagamento** O recebimento dos valores, nos moldes previstos em contrato, dependerá da apresentação dos respectivos documentos fiscais pelo contemplado, e os recursos deverão ser destinados conforme previsão contratual. Não serão permitidos pagamentos em moeda estrangeira. O Participante contemplado será o exclusivo responsável pela contratação e aquisição de todos os serviços e bens que estejam sob sua responsabilidade por contrato, bem como por efetuar os respectivos pagamentos e o recolhimento de impostos e demais encargos. O pagamento do valor será realizado em parcela única, condicionado à formalização do contrato e apresentação dos documentos fiscais cabíveis. O pagamento, quando respeitadas as condições indicadas neste Edital, será realizado nos seguintes prazos\*: – Notas fiscais ou recibos emitidos e enviados ao financeiro no período de 10 a 5 do mês – com vencimento no dia 18; – Notas fiscais ou recibos emitidos e enviados ao financeiro no período de 6 a 18 do mês – com vencimento no dia 5 do mês subsequente. \*Notas fiscais ou recibos emitidos entre os dias 18 e 31 não serão aceitos. Haverá prestação de contas ou outras responsabilidades do contemplado? Sim. A prestação de contas abrangerá a realização do projeto em conformidade com o pactuado entre as partes, o cumprimento das questões jurídicas, o atendimento do cronograma estabelecido, o uso adequado dos valores recebidos e a apresentação da documentação exigida em contrato. Os projetos deverão respeitar plenamente todas as obrigações jurídicas, legais e fiscais, ficando sujeitos a desclassificação em qualquer etapa do Edital em caso de inobservância. A Fundação

Itaú solicitará a apresentação de relatórios periódicos para acompanhamento da execução do projeto e dos pagamentos efetuados, bem como comprovantes fiscais e documentos relacionados às despesas e aos desembolsos efetuados pelo proponente, de acordo com os itens orçamentários aprovados. Haverá incidência de impostos nos valores destinados aos projetos contemplados? Sim, os valores divulgados e destinados ao contemplado serão sempre valores brutos. A tributação irá variar de acordo com o perfil do contemplado, conforme abaixo: Pessoa física – o valor pago à pessoa física sofrerá retenção de Imposto de Renda (IR) de acordo com a tabela progressiva, aplicado pela fonte pagadora (Fundação Itaú) em observação à legislação vigente no ano do pagamento. Pessoa jurídica – o valor pago à pessoa jurídica não sofrerá retenção de tributo federal na fonte pagadora a partir da emissão de recibo. Porém, se o proponente optar pela emissão de nota fiscal, poderá estar sujeito a retenção de Imposto sobre Serviço (ISS) pela fonte pagadora, em observação à legislação vigente da prefeitura de São Paulo. A Fundação Itaú não terá nenhuma responsabilidade pela incidência e pelo recolhimento dos impostos que tenham como contribuinte o Participante contemplado, sendo pessoa física ou jurídica. Contrapartidas A participação neste Edital e a anuênci a formulário de inscrição significam concordância com as contrapartidas que serão estabelecidas em contrato, como as citadas a seguir, sem pagamentos adicionais além dos já recebidos por meio deste Edital: a) Menção à Fundação Itaú como apoiadora em todas as divulgações e no uso do projeto em forma de logomarca, bem como menções em todas as ações e mídias de divulgação e uso do projeto; b) Autorização à Fundação Itaú – por meio de licenciamento – de direitos, não exclusivos, de utilização, divulgação, exibição, armazenagem, veiculação e difusão dos projetos contemplados e de seus resultados em diversos formatos, mídias, meios, suportes e modalidades; c) Compartilhamento de resultados com a Fundação Itaú. Os direitos da Fundação Itaú serão sempre não exclusivos e não irão prejudicar a utilização e a exploração, pelo Participante contemplado, de sua própria obra, de acordo com as condições previstas em contrato.

## **11. Condições de tratamento de dados**

A Fundação Itaú compromete-se a não utilizar, por si ou por terceiros, as informações fornecidas no formulário de inscrição pelos participantes, para fins diversos daqueles mencionados neste Edital. A Fundação Itaú poderá tratar dados pessoais presentes nos Formulários de inscrição e(ou) demais documentos fornecidos pelos inscritos, em razão da sua participação e do seu monitoramento do Edital e quando necessário para cumprir as obrigações legais aplicáveis, entre outros, para fins de exercício regular de direito em processo judicial, administrativo ou arbitral. A Fundação Itaú toma todas as providências técnicas e organizacionais para proteger os dados pessoais dos participantes contra perda, uso não autorizado ou outros abusos, garantindo o nível de segurança compatível com as atividades realizadas pela Fundação Itaú. Os dados pessoais serão armazenados em um ambiente operacional seguro e não acessível ao público. Nós nos esforçamos para proteger os dados pessoais que armazenamos, a partir de medidas, como: (I) adoção de um programa de governança e privacidade; (II) criação de equipe responsável pela proteção de dados pessoais e designação de encarregado pela proteção de dados; (III) adição dos princípios de “Privacy by Design” e “Privacy by Default” durante a execução das atividades; (IV) criação de políticas de confidencialidade com os colaboradores, (V) restrições de acesso aos dados com base nos princípios “need-to-know” e “least privilege”, dentre outras. Apesar disso, infelizmente não podemos garantir total segurança. O uso não autorizado de contas, falha de hardware ou software e outros fatores podem comprometer a segurança de dados pessoais a qualquer momento; por isso, ajude-nos a manter um ambiente seguro para todos. Os dados pessoais coletados em razão deste Edital poderão ser compartilhados com fornecedores e parceiros da Fundação Itaú e do conglomerado Itaú Unibanco com quem estes tenham relação em razão deste Edital. Além de adotar boas práticas de segurança em relação aos dados pessoais dos participantes, a Fundação Itaú se coloca à disposição para ajudá-los, caso identifique ou tome conhecimento de algo que comprometa a segurança de dados pessoais. Nesse caso, por favor entre em contato conosco pelos canais de atendimento indicados ao final deste Edital.

## **12. Como entrar em contato**

As dúvidas sobre este Edital serão respondidas somente pelo e-mail iaparaeducacao@fundacaoitau.org.br. Casos omissos deste Edital serão resolvidos pela Fundação Itaú, cuja decisão será soberana, e contra ela não caberá qualquer recurso. A Fundação Itaú poderá, a qualquer momento, desclassificar, suspender ou cancelar a inscrição de qualquer projeto, caso seja verificado fraude ou desatendimento de qualquer exigência deste Edital. Eventuais alterações neste Edital poderão ser realizadas a exclusivo critério da Fundação Itaú e serão divulgadas nos mesmos meios de comunicação das ações do Edital, passando a vigorar imediatamente após a publicação. Este Edital é regido de acordo com a legislação brasileira. Quaisquer disputas ou controvérsias oriundas de quaisquer atos praticados pelos participantes, inclusive com relação ao descumprimento do Edital ou pela violação de direitos da Fundação Itaú, de outros participantes, e/ou de terceiros, inclusive de propriedade intelectual, de sigilo e de personalidade, deverão ser processadas na comarca da capital do estado de São Paulo.