Gustavo R. Almeida

Framework Multicritério para Seleção de Modelos de Linguagem em Contextos Educacionais: Uma Análise de Desempenho, Custo e Eficiência Energética

Brasil

Julho de 2026

Gustavo R. Almeida

Framework Multicritério para Seleção de Modelos de Linguagem em Contextos Educacionais: Uma Análise de Desempenho, Custo e Eficiência Energética

Dissertação apresentada ao Programa de Pós-Graduação em Ciência e Tecnologia da Computação da Universidade Federal de Itajubá

Orientador: Adler D. Souza
Coorientador: Isabela N. Drummond

Brasil
Julho de 2026

Dedico este trabalho a todas as pessoas que
acreditam no poder da educação e dedicam seus esforços
para torná-la acessível a todos.

Agradecimentos

Agradeço, primeiramente, à minha futura esposa, Letícia Fernandes, por seu incentivo, por estar sempre ao meu lado e por nunca me deixar desistir.

Agradeço aos meus pais e às minhas irmãs, que sempre estiveram ao meu lado.

Agradeço aos meus orientadores, Professor Dr. Adler Diniz e Professora Dra. Isabela Drummond, por toda a orientação e parceria, acompanhando-me em todos os passos deste projeto.

Agradeço a todos os meus animais, que fazem parte da minha vida e a tornam mais feliz e tranquila.

Por fim, agradeço aos meus fiéis amigos, que estão sempre ao meu lado em todas as ocasiões.

"Tu és forte, tu és grande
Dentre os grandes és o primeiro"

Resumo

A adoção de modelos de inteligência artificial em Ambientes Virtuais de Aprendizagem (AVAs) exige equilibrar qualidade das respostas, latência, custo computacional e consumo energético, especialmente em cenários de uso contínuo e em larga escala. Este trabalho propõe um *framework* de decisão multicritério para a seleção entre *Small Language Models* (SLMs) e *Large Language Models* (LLMs) em contextos educacionais, modelando a escolha como um problema de aproximação de perfis (*Profile Matching*) baseado na distância euclidiana entre o perfil-alvo do cenário e o desempenho empírico normalizado de cada modelo. Sob a lógica da *Design Science Research* (DSR), foram avaliados 21 modelos (13 base e 8 com ajuste fino), variando de 1 a 70 bilhões de parâmetros, em 240 questões de múltipla escolha (120 originais e 120 do ENEM), distribuídas em oito disciplinas e três níveis de dificuldade. Os resultados indicam que os LLMs apresentam maior acurácia, com vantagem clara em questões difíceis, enquanto os SLMs não treinados são substancialmente mais rápidos (cerca de 8,56 vezes menor latência) e energeticamente mais eficientes (cerca de 4,83 vezes menos energia). O ajuste fino, nas condições avaliadas, elevou os custos operacionais sem ganho consistente de acurácia. Nos dez cenários simulados, o *framework* selecionou oito modelos distintos, evidenciando que a melhor escolha depende do contexto e que a seleção baseada em perfis supera a maximização de métricas isoladas.

Palavras-chave: inteligência artificial. ambientes virtuais de aprendizagem. modelos de linguagem. seleção multicritério. profile matching.

Abstract

The adoption of artificial intelligence models in Virtual Learning Environments (VLEs) requires a balance among response quality, latency, computational cost, and energy consumption, especially in continuous and large-scale usage scenarios. This study proposes a multicriteria decision *framework* for selecting between *Small Language Models* (SLMs) and *Large Language Models* (LLMs) in educational contexts, modeling the choice as a *Profile Matching* problem based on the Euclidean distance between the scenario's target profile and the normalized empirical performance of each model. Following the logic of *Design Science Research* (DSR), 21 models (13 base models and 8 fine-tuned models) were evaluated, ranging from 1 to 70 billion parameters, across 240 multiple-choice questions (120 original and 120 from ENEM), distributed among eight subjects and three difficulty levels. The results indicate that LLMs achieve higher accuracy, with a clear advantage on difficult questions, whereas untuned SLMs are substantially faster (approximately 8.56 times lower latency) and more energy-efficient (approximately 4.83 times less energy). Fine-tuning, under the conditions evaluated, increased operational costs without a consistent gain in accuracy. Across the ten simulated scenarios, the *framework* selected eight distinct models, demonstrating that the best choice depends on the context and that profile-based selection outperforms the maximization of isolated metrics.

Keywords: artificial intelligence. virtual learning environments. language models. multi-criteria selection. profile matching.

Lista de ilustrações

Figura 1 – *Design Science Research* da pesquisa. 40
Figura 2 – Etapas de estrutura experimental e demonstração do *framework*. 43
Figura 3 – Estrutura geral dos procedimentos experimentais da pesquisa. 53
Figura 4 – Visão geral do processo de ajuste fino supervisionado aplicado aos SLMs com dados educacionais. 60
Figura 5 – Visão geral da fase de análise. 66
Figura 6 – Acurácia por Modelo em Questões Não Originais 76
Figura 7 – Acurácia por Modelo em Questões Originais 77
Figura 8 – Latência por Modelo em Questões Não Originais 79
Figura 9 – Latência por Modelo em Questões Originais 79
Figura 10 – Custo Computacional por Modelo em Questões Não Originais 81
Figura 11 – Custo Computacional por Modelo em Questões Originais 81
Figura 12 – Consumo Energético Médio em joules por Modelo em Questões Não Originais 83
Figura 13 – Consumo Energético Médio em joules por Modelo em Questões Originais 83
Figura 14 – Potência média demandada em watts por Modelo em Questões Não Originais 85
Figura 15 – Potência média demandada em watts por Modelo em Questões Originais 85
Figura 16 – Potência máxima demandada em watts por Modelo em Questões Não Originais 87
Figura 17 – Potência máxima demandada em watts por Modelo em Questões Originais 87
Figura 18 – log(Tokens) vs parâmetros (B) 89
Figura 19 – log(Energia Total J) vs parâmetros (B) 90

Lista de tabelas

Tabela 1 – Valores de temperatura e seus efeitos na geração de texto. 29
Tabela 2 – Valores de *top-p* e seus efeitos na geração de texto. 30
Tabela 3 – Valores de *top-k* e seus efeitos na geração de texto. 30
Tabela 4 – Valores de *min-p* e seus efeitos na geração de texto. 31
Tabela 5 – Cenários de uso e respectivas configurações de busca no *framework* . . 50
Tabela 6 – Estrutura da tabela Perguntas 52
Tabela 7 – Estrutura da tabela Respostas 54
Tabela 8 – Estrutura da tabela Energia 54
Tabela 9 – Modelos de linguagem selecionados para experimentação 56
Tabela 10 – Relação entre hipóteses, variáveis e métodos estatísticos. 70
Tabela 11 – Acurácia por modelo em questões 76
Tabela 12 – Tempo médio de resposta dos modelos em questões 78
Tabela 13 – Custo Computacional por modelo em questões 80
Tabela 14 – Consumo energético médio em joules por modelo em questões 82
Tabela 15 – Potência média demandada em watts por modelo em questões 84
Tabela 16 – Potência máxima em watts por modelo em questões 86
Tabela 17 – Modelo selecionado e distância euclidiana por cenário 103

Lista de abreviaturas e siglas

IA	Inteligência Artificial
AVAs	Ambientes Virtuais de Aprendizagem
LLMs	Large Language Models
SLMs	Small Language Models
PLN	Processamento de Linguagem Natural
TICs	Tecnologias da Informação e Comunicação
MCDA	Multiple Criteria Decision Analysis
MCDM	Multiple Criteria Decision Making

Sumário

1 INTRODUÇÃO 13

2 REFERENCIAL TEÓRICO 15
2.1 Inteligência Artificial na Educação 15
2.2 Processamento de Linguagem Natural (PLN) e Educação 16
2.3 Small Language Models (SLMs) e Large Language Models (LLMs) 17
2.4 Ambientes Virtuais de Aprendizagem e Personalização 20
2.5 Benchmarks Educacionais e Conjunto de Dados 21
2.6 Decisão Multicritério (MCDA/MCDM) 22
2.7 Design Science Research (DSR) 23
2.8 *Trade-offs* entre SLMs e LLMs 24
2.9 Custo Energético e Sustentabilidade 24
2.10 Green AI, Custo e Energia em PLN e LLMs 26
2.11 Custo e Energia em Modelos de Linguagem 26
2.12 Ajuste Fino 27
2.13 Engenharia de Prompt 29
2.14 Desafios e Limitações 31

3 TRABALHOS RELACIONADOS 34

4 METODOLOGIA 39
4.1 Delineamento da pesquisa e lógica do artefato 39
4.2 Estrutura Experimental e Demonstração do *Framework* 41
4.3 Bases de dados e cenários de demonstração 42
4.4 Métricas 44
4.4.1 Métricas de Benefício 44
4.4.2 Métricas de Custo 44
4.4.3 Métricas de Avaliação 45
4.5 Função de Utilidade e Regras de Decisão 46
4.5.1 Definição das Faixas de Utilização 46
4.5.2 Modelagem Matemática de Aproximação de Perfil 48
4.6 Hipóteses e Variáveis 49
4.6.1 Dimensão 1: Desempenho por Complexidade 49
4.6.2 Dimensão 2: Eficiência Computacional 51
4.6.3 Dimensão 3: *Trade-offs* entre Desempenho e Custo 51
4.6.4 Dimensão 4: Adequação Contextual e Decisão 51

4.7 Procedimentos experimentais . . . 52
4.7.1 Fase de Preparação . . . 52
4.7.1.1 Implementação da Aplicação . . . 52
4.7.1.2 Modelos de linguagem . . . 55
4.7.1.3 Ajuste fino dos Modelos . . . 59
4.7.1.4 Engenharia de prompts e inferência . . . 62
4.7.2 Fase de Intervenção . . . 63
4.7.3 Fase de Análise . . . 65
4.7.3.1 Estrutura dos dados experimentais . . . 67
4.7.3.2 Variáveis analisadas . . . 67
4.7.3.3 Estratégia de análise estatística . . . 68
4.7.3.4 Análise das hipóteses . . . 70
4.7.3.5 Dimensão 1: Acurácia por Complexidade . . . 70
4.7.3.6 Dimensão 2: Eficiência Computacional . . . 71
4.7.3.7 Dimensão 3: *Trade-offs* entre Desempenho e Custo . . . 72
4.7.3.8 Dimensão 4: Adequação Contextual e Decisão . . . 73
4.7.3.9 Síntese da etapa analítica . . . 74

5 RESULTADOS E DISCUSSÃO . . . 75
5.1 Análise Experimental dos Modelos . . . 75
5.1.1 Análise Descritiva da Acurácia por Modelo . . . 75
5.1.2 Análise Descritiva da Latência por Modelo . . . 77
5.1.3 Análise Descritiva do Custo Computacional por Modelo . . . 78
5.1.4 Análise Descritiva do Consumo Energético por Modelo . . . 80
5.1.4.1 Consumo energético . . . 80
5.1.4.2 Potência Média em watts . . . 83
5.1.4.3 Potência Máxima em watts . . . 84
5.2 Testes Estatísticos . . . 86
5.2.1 Dimensão 1: Acurácia por Complexidade . . . 87
5.2.2 Dimensão 2: Eficiência Computacional e Energética . . . 88
5.2.3 Dimensão 3: *Trade-offs* e Adequação Contextual . . . 90
5.3 Demonstração do Framework . . . 92
5.3.1 Cenário 1: Tutoria em Tempo Real . . . 92
5.3.2 Cenário 2: Correção em Lote (*Batch*) . . . 93
5.3.3 Cenário 3: Monitoria Escalável para Grande Volume de Alunos . . . 94
5.3.4 Cenário 4: Raciocínio Lógico Complexo (Sem Restrição de Hardware) . . . 95
5.3.5 Cenário 5: Quiz Gamificado Mobile (Leve e Ágil) . . . 96
5.3.6 Cenário 6: Simulado ENEM com Feedback Explicativo . . . 97
5.3.7 Cenário 7: Sala de Aula Premium Multimodal . . . 98
5.3.8 Cenário 8: Escola com Recursos Limitados e Conectividade Instável . . . 99

5.3.9 Cenário 9: Personalização Adaptativa de Trilhas de Aprendizagem 100
5.3.10 Cenário 10: *Baseline* Equilibrado para AVA Geral 102
5.3.11 Síntese Geral e Hipóteses 102
5.4 Considerações Finais 105

6 CONCLUSÃO 107

7 USO DE INTELIGÊNCIA ARTIFICIAL 110

REFERÊNCIAS 111

1 Introdução

O avanço acelerado da Inteligência Artificial (IA) vem transformando de maneira profunda diversos setores da sociedade, com impactos particularmente relevantes no campo da educação (MELISA et al., 2025). Nesse contexto, a adoção de recursos computacionais inteligentes em Ambientes Virtuais de Aprendizagem (AVAs) tem ampliado as possibilidades de apoio ao processo de ensino-aprendizagem, favorecendo novas formas de interação, personalização e acompanhamento pedagógico por meio de softwares e sistemas acessíveis em computadores, celulares e *tablets* (LIMA et al., 2017).

Entre as tecnologias que mais têm impulsionado o desenvolvimento de soluções educacionais baseadas em IA, destacam-se os Large Language Models (LLMs) e os Small Language Models (SLMs), capazes de oferecer experiências de aprendizagem mais dinâmicas e adaptativas. Esses modelos podem emular interações humanas e responder de forma interativa às dúvidas dos estudantes, promovendo uma experiência de diálogo mais natural e contextualizada (CHOI; KIM, 2024).

Quando integrados aos AVAs, esses modelos podem contribuir para a construção de ambientes mais responsivos, adaptativos e engajadores, ajustando-se ao ritmo, às preferências e às necessidades individuais de aprendizagem (MOREIRA; LIMA, 2023). Além disso, os educadores também são beneficiados neste processo através da automatização de tarefas como avaliação e geração de conteúdos, o que possibilita maior foco em atividades pedagógicas estratégicas (VEMULA, 2024).

Evidências empíricas indicam que o uso de IA em Ambientes Virtuais de Aprendizagem pode impactar diretamente a retenção, o engajamento, a satisfação e o desempenho acadêmico dos estudantes, com reflexos positivos nas taxas de conclusão, retenção de conhecimento, motivação, confiança, produtividade, tempo e frequência de estudo (REZENDE; DINIZ; DRUMMOND, 2025). Entretanto, a adoção desses modelos em contextos educacionais não deve ser analisada apenas sob a perspectiva da qualidade das respostas. A escolha entre SLMs e LLMs também envolve *trade-offs* relacionados a custo computacional, latência, escalabilidade e consumo energético, especialmente em cenários de uso contínuo e em larga escala.

Nesse cenário, os SLMs têm ganhado destaque por apresentarem vantagens importantes, como menor custo computacional, maior previsibilidade e desempenho competitivo em tarefas específicas, o que os torna promissores para aplicações educacionais mais acessíveis e escaláveis (REZENDE; DINIZ; DRUMMOND, 2025). Ainda assim, a literatura aponta uma lacuna de estudos sistemáticos que investiguem, de forma comparativa e contextualizada, o comportamento de SLMs e LLMs em diferentes cenários educacionais, considerando

simultaneamente métricas de desempenho, custo e sustentabilidade.

Essa questão se torna ainda mais relevante diante das preocupações associadas ao impacto ambiental da IA. O treinamento e a inferência de modelos de linguagem de grande porte demandam elevado consumo de energia e recursos hídricos, ampliando o debate em torno da sustentabilidade computacional e do conceito de *Green AI* (ECCARD; CAVALCANTE et al., 2025). Assim, além de avaliar a acurácia dos modelos, torna-se necessário considerar sua eficiência energética, especialmente em contextos em que a adoção em larga escala pode intensificar custos ambientais e operacionais.

Diante desse cenário, o objetivo desta pesquisa é propor um *framework* de decisão multicritério para seleção entre SLMs e LLMs em contextos educacionais, considerando não apenas o desempenho na acurácia das respostas, mas também aspectos como tempo de resposta, consumo de tokens e consumo energético. A proposta busca oferecer um mecanismo estruturado para apoiar a escolha do modelo mais adequado conforme o cenário de uso, reconhecendo que não existe um modelo universalmente superior, mas sim modelos mais ou menos apropriados de acordo com a complexidade da tarefa, a disciplina avaliada e o nível de dificuldade das questões.

A investigação adota uma experimentação computacional controlada, na qual diferentes SLMs e LLMs, treinados e não treinados, são avaliados por meio de conjuntos de questões de múltipla escolha originais e não originais, fundamentadas em disciplinas do ensino médio, a saber: História, Geografia, Matemática, Português, Inglês, Biologia, Física e Química. As questões são classificadas nos níveis de dificuldade fácil, médio e difícil, permitindo observar o comportamento dos modelos em diferentes graus de complexidade cognitiva. Cada modelo é testado sobre o mesmo conjunto de questões, com o objetivo de reduzir variabilidade e possibilitar uma análise comparativa robusta.

Como resultado, espera-se gerar evidências técnicas e quantitativas capazes de subsidiar a construção e a demonstração do *framework* proposto, contribuindo para a compreensão dos *trade-offs* entre desempenho, custo e sustentabilidade na aplicação de modelos de linguagem em educação. Dessa forma, esta pesquisa busca avançar além da comparação entre modelos, propondo um instrumento de apoio à decisão que considere, de forma integrada, os requisitos pedagógicos, computacionais e ambientais das aplicações educacionais baseadas em IA.

Este trabalho está organizado da seguinte forma: na Seção 2, apresenta-se o referencial teórico sobre os temas abordados; na Seção 3, são discutidos os trabalhos relacionados; na Seção 4, detalha-se a metodologia utilizada; na Seção 5, são apresentados os resultados e discussões; na Seção 6, são discutidas as conclusões e implicações da pesquisa; e, por fim, na Seção 7, são apresentadas as utilizações de Inteligência Artificial neste trabalho.

2 Referencial Teórico

2.1 Inteligência Artificial na Educação

O avanço contínuo das tecnologias digitais tem provocado transformações profundas no cotidiano humano, alterando de maneira significativa a forma como vivemos, trabalhamos e interagimos. Nesse contexto, a Inteligência Artificial (IA) desponta como uma das principais forças impulsionadoras dessa mudança, assumindo papel central em diferentes esferas da sociedade, influenciando nossas rotinas diárias, desde a maneira como consumimos informação até os modos de comunicação e tomada de decisão (OH et al., 2024).

Ao longo da última década, o campo da IA expandiu-se em ritmo acelerado, demonstrando utilidade em setores diversos, como: criação de conteúdo digital; desenvolvimento pessoal; atendimento ao cliente; tradução automática de idiomas; entretenimento; serviços profissionais especializados; análise preditiva; e marketing. Essa multiplicidade de aplicações evidencia a versatilidade da IA e cria um terreno fértil para sua incorporação no cenário educacional, na qual as demandas por inovação pedagógica, acessibilidade e personalização são crescentes (YADAV; ARORA; KASHYAP, 2025).

No domínio da educação, a IA tem se consolidado como aliada estratégica para apoiar e ampliar o processo de ensino-aprendizagem. Entre as principais possibilidades de uso, destacam-se: a criação de ambientes virtuais de aprendizagem mais interativos e dinâmicos; a oferta de suporte contínuo aos alunos, com disponibilidade 24 horas por dia, sete dias por semana; a personalização das trilhas de aprendizagem, adaptando conteúdos e métodos às necessidades individuais; a recomendação e curadoria de recursos educacionais de acordo com o perfil do estudante; a automatização de processos administrativos, como admissões, matrículas e planejamento de cursos; a preparação direcionada para exames; a avaliação automatizada de atividades; o atendimento a necessidades específicas de estudantes com deficiências ou dificuldades de aprendizagem; e o aconselhamento acadêmico e profissional orientado por dados (YADAV; ARORA; KASHYAP, 2025).

Na prática, essas possibilidades já se materializam em uma ampla gama de plataformas e ferramentas educacionais. Exemplos incluem os tutores virtuais, que fornecem feedback instantâneo e acompanhamento individualizado; os assistentes de voz, como Google Assistente, Siri e Cortana, que auxiliam em tarefas de busca e organização; os sistemas de geração automática de conteúdo educacional; os tradutores automáticos aplicados em apresentações e materiais didáticos multilíngues; os cursos online massivos (MOOCs), como Udemy, Coursera, Khan Academy e edX, que incorporam mecanismos de recomendação; os sistemas de avaliação automática baseados em processamento de

linguagem natural; os jogos educacionais com elementos de gamificação; e, de forma mais sofisticada, os Sistemas Tutoriais Inteligentes (ITS – Intelligent Tutoring Systems) ou instruções assistidas por computador inteligente (ICAI – Intelligent Computer-Assisted Instruction) (SENANAYAKE; KARUNANAYAKA; EKANAYAKE, 2024).

Assim, a presença da IA na educação não apenas amplia o acesso ao conhecimento, mas também potencializa a qualidade e a eficiência dos processos de ensino. Ao oferecer experiências de aprendizagem mais dinâmicas, adaptativas e centradas no aluno, a IA contribui para o desenvolvimento de competências cognitivas avançadas, como pensamento crítico, resolução de problemas e capacidade analítica. Além disso, favorece uma experiência educacional mais agradável e engajadora, capaz de se ajustar ao ritmo e às preferências individuais, promovendo maior retenção do conhecimento e motivação dos estudantes (YADAV; ARORA; KASHYAP, 2025; ALI et al., 2024; SENANAYAKE; KARUNANAYAKA; EKANAYAKE, 2024).

Embora a IA possibilite a ampliação das possibilidades educacionais, a adoção em AVAs exige critérios de escolha que abrangem não apenas a qualidade das respostas, mas também custo computacional, tempo de resposta e consumo energético, especialmente em cenários de uso em larga escala. Nesse sentido, torna-se necessário um mecanismo de decisão capaz de orientar a seleção do modelo mais adequado à aquele contexto educacional.

2.2 Processamento de Linguagem Natural (PLN) e Educação

Os Modelos de Processamento de Linguagem Natural (PLN) têm como propósito central possibilitar que sistemas computacionais compreendam, interpretem, processem e até mesmo produzam a linguagem humana de forma cada vez mais natural e fluida. Essa capacidade constitui um dos pilares fundamentais para o avanço da Inteligência Artificial, uma vez que torna viável a criação de aplicações que interagem com os usuários de maneira mais próxima da comunicação humana, reduzindo barreiras tecnológicas e promovendo experiências mais intuitivas (YADAV; ARORA; KASHYAP, 2025).

O progresso recente na área de PLN está fortemente associado a três fatores convergentes. Em primeiro lugar, o desenvolvimento e a maturidade das técnicas de aprendizado profundo (deep learning), que permitiram a construção de arquiteturas de redes neurais mais robustas, como os transformadores. Em segundo lugar, o acesso a enormes volumes de dados textuais disponíveis em escala global — provenientes da internet, da literatura científica, de mídias digitais, jornais e redes sociais — os quais se tornaram a base para treinar modelos linguísticos em larga escala. Por fim, os avanços significativos na capacidade de processamento computacional, possibilitados por GPUs, TPUs e infraestrutura em nuvem, que permitiram não apenas treinar modelos complexos, mas também disponibilizá-los em tempo real para uso cotidiano. Esses elementos

combinados viabilizaram o surgimento de modelos altamente sofisticados, capazes de lidar com tarefas linguísticas complexas, como tradução automática, sumarização de textos, análise de sentimentos, geração de conteúdo e diálogos contextuais (YADAV; ARORA; KASHYAP, 2025).

A partir de meados da década de 2010, com a popularização do aprendizado de máquina aplicado ao PLN e o lançamento de APIs de alto nível acessíveis a desenvolvedores, consolidou-se um novo paradigma no design de software: os assistentes virtuais. Esses sistemas, que se manifestam por meio de interfaces de texto ou de voz, utilizam modelos de linguagem para mediar a interação homem-máquina de maneira mais natural. Exemplos conhecidos incluem Siri, Alexa, Google Assistente e Cortana, que oferecem desde comandos simples de organização até interações mais complexas baseadas em contexto. No campo educacional, esses assistentes têm ampliado seu papel, servindo como ferramentas de apoio personalizadas, capazes de responder dúvidas em tempo real, recomendar conteúdos e auxiliar na organização de rotinas de estudo (ALIER et al., 2025).

Portanto, o avanço dos modelos baseados em PLN representa um marco para a educação contemporânea. Ao permitir interações mais humanizadas e centradas no estudante, essas tecnologias reforçam o potencial de personalização, acessibilidade e escalabilidade dos processos de ensino-aprendizagem, criando condições para experiências formativas mais ricas e inclusivas. Ao mesmo tempo, sua adoção requer avaliação cuidadosa de desempenho, eficiência e adequação ao contexto de uso, especialmente em busca de apoio automatizado à aprendizagem, pois cada modelo tem sua arquitetura, envolvendo custos e capacidades distintas, necessitando de uma utilização contextualizada ao cenário.

2.3 Small Language Models (SLMs) e Large Language Models (LLMs)

Atrelados aos avanços no campo do Processamento de Linguagem Natural (PLN), emergem os Modelos de Linguagem Pequenos (SLMs – Small Language Models) e os Modelos de Linguagem Grandes (LLMs – Large Language Models). Essas duas categorias de modelos representam abordagens complementares no desenvolvimento de sistemas inteligentes, distinguindo-se pela escala, pelos requisitos computacionais, pelas aplicações práticas possíveis e pelo cenário de uso.

A introdução da arquitetura Transformer, em 2017, revolucionou o campo da IA ao permitir que redes neurais processassem sequências de texto de forma contextual, utilizando mecanismos de autoatenção para capturar relações semânticas entre palavras em diferentes posições, sendo fundamental para a evolução dos modelos de linguagem. Associado a essa arquitetura, o conceito das “leis de escala” mostrou que, ao aumentar o número de parâmetros, a quantidade de dados de treinamento e a capacidade computacional

disponível, os modelos tendem a apresentar ganhos consistentes de desempenho em diversas tarefas linguísticas (LI et al., 2024).

Essa constatação impulsionou a rápida ascensão dos LLMs. Os LLMs são modelos que caracterizam-se por seu tamanho massivo, com bilhões ou até centenas de bilhões de parâmetros, sendo treinados em conjuntos de dados textuais amplos e heterogêneos, coletados em escala global. Como resultado, demonstram desempenho notável em tarefas complexas, como geração de texto coerente, tradução automática, sumarização de documentos extensos, raciocínio lógico, análise de sentimentos e resposta a perguntas em múltiplos domínios, isso tudo mesmo possuindo pouca instrução explícita (few-shot learning) (CHOI; KIM, 2024; HUANG et al., 2022).

Entre os principais LLMs representativos atualmente disponíveis, destacam-se (CHANDRASEKARAN; NAIR, 2024):

- **GPT** (OpenAI): Ferramenta de geração de texto amplamente utilizada em chatbots, tradução e criação de conteúdo (<https://openai.com/>).
- **Gemini** (Google): Focado em recuperação de informações e geração de código, com aplicações em assistência à escrita e análise de dados (<https://gemini.google/br/about>).
- **Claude** (Anthropic): Desenvolvido com ênfase em segurança e mitigação de vieses, aplicado em educação, revisão de código e suporte corporativo (<https://www.anthropic.com/>).
- **LLaMA** (Meta): Voltado para pesquisa e aplicações acadêmicas, com versões adaptadas para diferentes demandas de escala (<https://www.llama.com/>).
- **Vicuna** (LMSYS): Projetado para interpretabilidade e explicações, sendo útil em contextos científicos e pedagógicos (<https://www.lmsys.org/>).
- **Stable Beluga** (Stability AI): Combina geração de texto com produção de imagens, voltado para marketing e áreas criativas (<https://stability.ai/>).

Entretanto, atualmente esses modelos não representam uma superioridade universal, devido ao fato de a adequação de cada modelo depender do contexto de uso. O desempenho dos modelos não pode ser tratado apenas com a taxa de acurácia deles, e sim juntamente com seu custo, latência e energia. Com base nisso, mais recentemente, observa-se uma mudança no foco das pesquisas: da simples ampliação de escala para a busca por maior eficiência, qualidade de dados e acessibilidade. Essa tendência impulsionou o desenvolvimento dos SLMs, que se apresentam como modelos menores, mais leves e energeticamente eficientes, reformulando a lógica das leis de escala (LI et al., 2024).

Os SLMs geralmente possuem de milhões a poucos bilhões de parâmetros, sendo treinados em conjuntos de dados mais compactos e frequentemente especializados em domínios específicos. Essa configuração resulta em benefícios relevantes: menor custo de treinamento e operação, maior velocidade de inferência, menor consumo energético e a possibilidade de execução em dispositivos locais com recursos limitados, como smartphones, tablets ou servidores de baixo custo (ZHANG et al., 2024). Além disso, modelos menores favorecem cenários em que a privacidade e a soberania de dados são críticas, já que podem ser executados localmente sem necessidade de envio de informações sensíveis para a nuvem.

Entre os exemplos representativos recentes de SLMs notáveis estão: o **Phi-3** (Microsoft), que se destaca pela eficiência em tarefas educacionais e de raciocínio; o **LLaMA 3** em sua versão otimizada (Meta), projetado para rodar em ambientes com restrições de hardware; e o **Mistral Small** (Mistral AI), conhecido por equilibrar velocidade e precisão em contextos práticos.

Quando aplicados a interfaces de conversação — como chatbots, tutores virtuais ou sistemas de apoio ao ensino — tanto LLMs quanto SLMs podem ser integrados em assistentes inteligentes. Esses assistentes combinam a capacidade de compreensão e geração de linguagem natural com o acesso a bases de conhecimento externas, possibilitando respostas mais contextualizadas, explicativas, citáveis e relevantes (ALIER et al., 2025).

No contexto educacional, a distinção entre LLMs e SLMs torna-se estratégica. Os LLMs oferecem maior capacidade de raciocínio e de geração de conteúdo sofisticado, o que os torna valiosos em cenários de suporte avançado, pesquisa acadêmica ou produção de materiais educativos complexos. Já os SLMs apresentam vantagens em ambientes de ensino com restrições de infraestrutura, em que se exige privacidade, rapidez e personalização local, como em escolas públicas, ambientes corporativos ou dispositivos móveis dos alunos, além de representar possíveis vantagens relacionadas a consumo energético, eficiência e latência. Ambos, no entanto, compartilham o potencial de revolucionar a educação ao oferecer feedback imediato, identificar erros conceituais, adaptar-se ao nível de cada estudante e promover trilhas de aprendizagem personalizadas (CHOI; KIM, 2024).

Com isso, a escolha entre SLMs e LLMs não deve ser entendida apenas como uma comparação de desempenho absoluto, mas como um problema de decisão multicritério. Isso ocorre porque a seleção do modelo mais adequado depende da ponderação conjunta de fatores como acurácia, tempo de resposta, custo computacional e consumo energético, os quais podem variar conforme a complexidade da tarefa e o cenário educacional em questão. Assim, mais do que identificar qual modelo apresenta melhor resultado em termos gerais, torna-se necessário definir qual modelo oferece o melhor equilíbrio entre desempenho e eficiência para cada contexto de uso, o que reforça a pertinência de um *framework* de apoio à decisão.

Assim, a coexistência e complementaridade entre SLMs e LLMs apontam para

um futuro em que diferentes modelos poderão ser selecionados e integrados de acordo com as necessidades específicas do ambiente educacional, equilibrando qualidade, custo e acessibilidade.

2.4 Ambientes Virtuais de Aprendizagem e Personalização

Os Ambientes Virtuais de Aprendizagem (AVAs) têm desempenhado um papel cada vez mais relevante, contribuindo significativamente para a evolução da educação mediada por tecnologia. Esses sistemas digitais têm como principal objetivo disponibilizar um espaço virtual estruturado para a construção do conhecimento, viabilizando a realização de atividades pedagógicas por meio das Tecnologias da Informação e Comunicação (TICs), por meio de uma personalização automática ou semi-automática. Nesse sentido, os AVAs funcionam como ecossistemas que integram conteúdos, atividades, interações sociais e recursos avaliativos, permitindo que o processo de ensino-aprendizagem ocorra de forma síncrona ou assíncrona, ultrapassando as limitações físicas da sala de aula tradicional (BARTELLE; MEDEIROS, 2024).

Nos últimos anos, a incorporação de assistentes virtuais e sistemas inteligentes nesses ambientes tem ampliado significativamente seu potencial pedagógico. Esses agentes digitais oferecem suporte imediato aos alunos, seja por meio da resolução de dúvidas em linguagem natural, da recomendação de recursos complementares ou do acompanhamento contínuo de desempenho. Um de seus principais diferenciais é a disponibilidade constante — 24 horas por dia, sete dias por semana —, o que permite ao estudante acessar informações e orientações de forma autônoma, no momento em que surge a necessidade. Esse aspecto favorece experiências de aprendizagem mais flexíveis, adaptadas ao ritmo e às circunstâncias individuais de cada aluno, reforçando a autonomia no processo de formação do aluno (BARTELLE; MEDEIROS, 2024).

Nesse contexto, o uso das técnicas de Inteligência Artificial (IA) nos AVAs impulsiona uma abordagem verdadeiramente centrada no estudante. A IA possibilita a personalização de conteúdos, trilhas de aprendizagem e modos de interação, considerando variáveis como estilo cognitivo, preferências de estudo, histórico de desempenho e necessidades específicas de cada indivíduo. Essa personalização favorece não apenas o aumento da motivação e do engajamento, mas também a melhoria da compreensão conceitual e a retenção de informações. Além disso, reduz as barreiras comuns ao ensino tradicional, como o ritmo padronizado e a falta de acompanhamento individualizado (ALI et al., 2024).

Outro aspecto relevante é a capacidade da IA de processar e analisar grandes volumes de dados educacionais gerados nas interações com os AVAs. Esse recurso permite identificar padrões de comportamento, dificuldades recorrentes e níveis de engajamento, oferecendo insights valiosos para os educadores. A partir dessas análises, é possível

implementar estratégias pedagógicas mais eficazes, criar intervenções proativas e promover ajustes contínuos nos métodos de ensino. Dessa forma, a inteligência artificial não apenas potencializa o suporte direto ao estudante, mas também se torna uma ferramenta estratégica para os professores e gestores educacionais, contribuindo para a melhoria global da qualidade do ensino e para a promoção do sucesso acadêmico (ALI et al., 2024).

Dessa forma, a Inteligência Artificial desempenha um papel relevante no fortalecimento da aprendizagem em Ambientes Virtuais de Aprendizagem (AVAs), ao oferecer suporte personalizado, dinâmico e potencialmente mais eficiente aos estudantes. Entretanto, para que esses benefícios sejam efetivamente alcançados, é necessário que o modelo adotado em cada ambiente seja selecionado de acordo com as demandas específicas daquele contexto. Nesse sentido, torna-se essencial considerar não apenas a capacidade de geração de respostas, mas também critérios como desempenho, custo computacional, tempo de resposta e eficiência energética, de modo a garantir que a escolha do modelo contribua efetivamente para melhores resultados educacionais.

2.5 Benchmarks Educacionais e Conjunto de Dados

A avaliação de modelos de linguagem em contextos educacionais requer a utilização de bases de dados confiáveis e representativas, que permitam mensurar a capacidade dos modelos em lidar com diferentes domínios disciplinares e níveis de dificuldade. Nesse sentido, benchmarks educacionais têm sido amplamente empregados como forma de padronizar comparações, fornecendo conjuntos de questões provenientes de exames reais, como avaliações nacionais e internacionais, ou bancos de dados específicos de disciplinas (FANG; OBERSKI; NGUYEN, 2024; CHEN et al., 2025).

Esses benchmarks geralmente incluem questões de múltipla escolha e discursivas, classificadas por disciplinas e níveis de dificuldade (fácil, médio, difícil), de modo a avaliar não apenas a capacidade de resposta direta, mas também o raciocínio contextual dos modelos. Além disso, novas propostas de avaliação vêm integrando princípios psicométricos para aproximar o desempenho dos modelos ao de estudantes humanos, o que aumenta a validade pedagógica das análises comparativas (FANG; OBERSKI; NGUYEN, 2024).

Assim, o uso de benchmarks educacionais possibilita a reprodução de cenários reais de avaliação escolar, oferecendo subsídios técnicos para a análise da adequação de SLMs e LLMs em diferentes contextos de ensino, como diferentes disciplinas e dificuldades, além de servir como base para uma decisão orientada a escolha de modelos.

2.6 Decisão Multicritério (MCDA/MCDM)

Os problemas de decisão em contextos reais, de maneira geral, devem ser decididos através da escolha entre diversas alternativas, que envolvem objetivos múltiplos, muitas vezes conflitantes, que precisam ser ponderados de forma estruturada e transparente. Nesse cenário, a Decisão Multicritério, também referida como *Multiple Criteria Decision Analysis* (MCDA) ou *Multiple Criteria Decision Making* (MCDM), constitui um campo metodológico dedicado a abordar cenários complexos de tomada de decisão que envolvem múltiplos critérios e que as dimensões devem ser consideradas simultaneamente. (MARTINS, 2025; BELTON; STEWART, 2012).

De forma geral, esta abordagem é especialmente valiosa em situações em que as decisões não podem ser avaliadas por meio de uma única perspectiva, uma vez que permite uma análise mais abrangente ao considerar critérios quantitativos e qualitativos, em situações nas quais não existe dominância absoluta entre as alternativas. Com isso, a decisão multicritério permite representar *trade-offs* entre atributos favorecendo decisões mais consistentes com os objetivos detalhados (MARTINS, 2025; BELTON; STEWART, 2012).

Entre os métodos clássicos da área, destaca-se o *Analytic Hierarchy Process* (AHP), proposto por (SAATY et al., 1980), amplamente utilizado para estruturar problemas decisórios em hierarquias de critérios e subcritérios, atribuindo pesos relativos por meio de comparações paritárias. Além do AHP, existe uma ampla variedade de abordagens compensatórias e não compensatórias, cada uma com pressupostos próprios sobre a agregação de preferências e da interpretação dos resultados. Essa diversidade metodológica evidencia que a escolha do método multicritério deve ser coerente com a natureza do problema, com a disponibilidade de dados e com os objetivos da decisão (BELTON; STEWART, 2012; MARTINS, 2025).

No contexto educacional e computacional, a aplicação da MCDA/MCDM torna-se essencial quando a seleção de uma alternativa depende de fatores técnicos e operacionais simultaneamente. No contexto desta pesquisa, a escolha entre diferentes modelos de linguagem, por exemplo, não pode ser reduzida apenas à acurácia das respostas, pois aspectos como latência, custo computacional, consumo energético e adequação ao ambiente de uso influenciam diretamente a viabilidade da solução. Dessa forma, a adoção de uma abordagem multicritério fornece suporte teórico para decisões mais equilibradas, permitindo a integração sistemática de critérios de natureza distinta em um mesmo processo avaliativo (KEENEY; RAIFFA, 1993; BELTON; STEWART, 2012; MARTINS, 2025).

Nesse sentido, a MCDA/MCDM oferece uma base conceitual apropriada para a seleção de modelos em Ambientes Virtuais de Aprendizagem, especialmente quando a decisão envolve o equilíbrio entre desempenho técnico e restrições práticas. Ao considerar

simultaneamente múltiplos critérios, essa metodologia contribui para a formulação de decisões mais justificáveis, rastreáveis e aderentes às necessidades específicas do contexto educacional.

2.7 Design Science Research (DSR)

A pesquisa em Sistemas de Informação e áreas correlatas frequentemente exige não apenas a compreensão de fenômenos existentes, mas também a proposição de soluções concretas para problemas práticos. Nesse cenário, a *Design Science Research* (DSR) constitui uma abordagem metodológica orientada à criação e à avaliação de artefatos destinados a resolver problemas relevantes em contextos organizacionais, sociais ou tecnológicos (HEVNER et al., 2008; PEFFERS et al., 2007).

A DSR combina rigor científico e relevância prática ao propor artefatos que podem assumir a forma de constructos, modelos, métodos, instâncias de sistemas ou *frameworks*. Nessa perspectiva, a pesquisa não se limita à descrição de um fenômeno, mas envolve a elaboração de uma solução inovadora, acompanhada de análise sistemática de sua utilidade, desempenho e aderência ao problema em estudo. Com isso, o conhecimento produzido possui dupla finalidade: contribuir para a teoria e, simultaneamente, atender a necessidades concretas (HEVNER et al., 2008).

A DSR é um processo metodológico composto por etapas inter-relacionadas, que incluem a identificação do problema, a definição dos objetivos da solução, o projeto e o desenvolvimento do artefato, sua demonstração, avaliação e comunicação dos resultados. Esse encadeamento metodológico é especialmente útil quando o objetivo da pesquisa é propor um mecanismo de apoio à decisão, uma vez que permite estruturar de forma clara tanto os requisitos da solução quanto os critérios de validação adotados (PEFFERS et al., 2007).

Dessa forma, a adoção da Design Science Research como método de construção do artefato reforça a natureza aplicada da pesquisa e fornece um caminho sistemático para a elaboração de soluções orientadas por evidências. Ao articular problema, proposta e avaliação, a DSR contribui para a produção de conhecimento útil e para o desenvolvimento de instrumentos (HEVNER et al., 2008; PEFFERS et al., 2007).

No contexto desta pesquisa, a DSR apresenta elevada aderência, pois o objetivo não é apenas analisar teoricamente modelos de linguagem, mas construir um artefato capaz de apoiar sua seleção em ambientes educacionais com base em múltiplos critérios. Nesse sentido, o *framework* proposto pode ser compreendido como um artefato de natureza metodológica e decisória, concebido para integrar métricas de desempenho, custo, energia e adequação contextual em um processo estruturado de avaliação.

Além disso, a escolha entre SLMs e LLMs, especialmente quando orientada por critérios educacionais e de sustentabilidade, não se resolve apenas por meio de análise teórica abstrata, mas exige uma solução que possa ser operacionalizada em cenários reais. Com isso, a DSR oferece um enquadramento metodológico adequado para a construção de um artefato que responda a essa demanda, ao mesmo tempo em que permite sua validação em termos de utilidade, consistência e efetividade.

2.8 *Trade-offs* entre SLMs e LLMs

A literatura recente tem apontado que a escolha entre Small Language Models (SLMs) e Large Language Models (LLMs) não deve se restringir apenas à acurácia das respostas, mas envolver também uma análise dos *trade-offs* associados. Enquanto os LLMs tendem a apresentar maior capacidade de raciocínio e abrangência de domínios, os SLMs oferecem maior eficiência, menor custo computacional e menor consumo energético, tornando-se atrativos em contextos práticos e educacionais (MARCO; RELLO; GONZALO, 2024; PECHER; SRBA; BIELIKOVA, 2024).

Os *trade-offs* envolvem, portanto, a busca por um equilíbrio entre qualidade da resposta, tempo de processamento e custo de energético. Estudos têm demonstrado que, em tarefas em diferentes cenários e contextos, SLMs podem alcançar desempenho comparável a LLMs, especialmente quando são especializados em domínios específicos. Já em tarefas de maior complexidade, os LLMs tendem a manter uma vantagem, embora com maior custo associado (MARCO; RELLO; GONZALO, 2024).

Dessa forma, não existe um modelo universalmente superior para todos os cenários, mas sim modelos mais ou menos adequados conforme o contexto de uso. A seleção entre SLMs e LLMs deve, portanto, considerar o equilíbrio entre acurácia, custo computacional, consumo energético e latência, reconhecendo que diferentes aplicações educacionais podem demandar diferentes prioridades. Nesse cenário, torna-se necessário um *framework* de apoio à decisão que permita identificar, de maneira sistemática, qual modelo melhor se ajusta a cada situação específica.

2.9 Custo Energético e Sustentabilidade

Um aspecto cada vez mais relevante na adoção de modelos de linguagem diz respeito ao seu impacto energético e ambiental. Modelos de grande porte (LLMs) exigem elevado consumo de energia, tanto em fase de treinamento quanto em inferência, o que levanta preocupações relacionadas à sustentabilidade da Inteligência Artificial (PODDAR et al., 2025; SAMSI et al., 2023).

A título de comparação, o treinamento de modelos de IA com bilhões de parâmetros

na técnica de aprendizado pode resultar em emissões de carbono comparáveis ao uso de cinco carros ao longo de sua vida útil. Esses fatores elevam as preocupações sobre o impacto ambiental da tecnologia, especialmente com sua expansão em escala (ECCARD; CAVALCANTE et al., 2025).

Um dos principais fatores que contribuem para o alto consumo energético dos treinamentos se dá pelo fato que os modelos requerem inúmeras iterações e ajustes de parâmetros, o que exige um grande poder computacional. Tomando como exemplo, o modelo GPT-3 da OpenAI, que possui 175 bilhões de parâmetros, exigiu um consumo de 1,287 MWh de eletricidade para seu treinamento completo, quantidade de energia equivalente ao consumo anual de cerca de 120 residências nos Estados Unidos (ECCARD; CAVALCANTE et al., 2025).

Além disso, a corrida tecnológica entre grandes empresas de tecnologia tem impulsionado o aumento do consumo de energia, devido a competição para desenvolver modelos de IA mais avançados e poderosos, requerindo infraestruturas cada vez mais robustas e data centers de maior escala (ECCARD; CAVALCANTE et al., 2025).

Nesse contexto, os SLMs emergem como alternativas mais sustentáveis, pois demandam menor quantidade de recursos computacionais, podem ser executados em dispositivos locais e reduzem a dependência de infraestrutura em nuvem. Estudos recentes têm buscado mensurar o custo energético da inferência em modelos de diferentes tamanhos, evidenciando a relação direta entre número de parâmetros, tempo de execução e gasto energético (PODDAR et al., 2025; FERNANDEZ et al., 2025a).

Além do aspecto ambiental, a redução do custo energético implica também em maior acessibilidade, viabilizando o uso de modelos em ambientes educacionais com infraestrutura limitada, como escolas públicas ou regiões com restrições tecnológicas. Assim, a análise do custo energético e da sustentabilidade não apenas reforça a justificativa para o uso de SLMs, como também amplia sua relevância em cenários de democratização do acesso à educação mediada por IA.

Nesse sentido, a questão energética não deve ser compreendida apenas como um aspecto de impacto ambiental, mas também como um critério relevante para a decisão computacional em cenários que envolvem múltiplas variáveis de avaliação. Ao incorporar o consumo energético como uma dimensão do processo decisório, o *framework* proposto possibilita que a sustentabilidade computacional seja considerada de forma explícita na seleção do modelo, permitindo equilibrar desempenho, custo e eficiência energética em contextos educacionais. Dessa maneira, a energia deixa de ser tratada como um efeito colateral da execução dos modelos e passa a compor, de forma objetiva, o conjunto de critérios que orientam a escolha mais adequada.

2.10 Green AI, Custo e Energia em PLN e LLMs

O crescimento acelerado dos modelos de aprendizado profundo e, em particular, dos modelos de linguagem em larga escala, intensificou debates acerca de seu custo computacional e de seu impacto ambiental. Em resposta a esse cenário, consolidou-se a agenda denominada *Green AI*, que propõe deslocar o foco exclusivo do desempenho preditivo para uma avaliação mais ampla, na qual a eficiência computacional, o consumo energético e a sustentabilidade também passam a ser considerados como critérios centrais de pesquisa e desenvolvimento (SCHWARTZ et al., 2020).

A proposta de Green AI emerge como contraponto à tendência de valorização irrestrita de modelos cada vez maiores, cujo treinamento e inferência demandam elevados volumes de energia, infraestrutura robusta e grande capacidade de processamento. Essa preocupação é particularmente relevante no Processamento de Linguagem Natural (PLN), área em que o uso de grandes corpora, arquiteturas profundas e treinamento em larga escala tem ampliado substancialmente a pegada energética dos sistemas (SCHWARTZ et al., 2020; STRUBELL; GANESH; MCCALLUM, 2019). Assim, a eficiência deixa de ser um atributo secundário e passa a compor uma dimensão essencial da avaliação de soluções baseadas em IA.

A preocupação com energia e sustentabilidade também motivou o desenvolvimento de benchmarks e iniciativas voltadas à medição padronizada de desempenho energético e de eficiência computacional. Entre elas, destacam-se os esforços da MLCommons, que propõem métricas e cenários comparáveis para avaliação de inferência e de consumo energético em sistemas de aprendizado de máquina (TSCHAND et al., 2025). Tais iniciativas reforçam a importância de considerar não apenas a qualidade das saídas produzidas pelos modelos, mas também os recursos necessários para sua execução.

No contexto deste trabalho, a discussão sobre Green AI é pertinente pois a seleção de modelos para uso em ambientes educacionais não deve ignorar o custo energético associado à sua operação. Modelos mais eficientes podem favorecer maior escalabilidade, menor dependência de infraestrutura e melhor adequação a cenários com restrições de recursos, ao mesmo tempo em que contribuem para uma adoção tecnologicamente responsável. Assim, a eficiência energética deixa de ser apenas uma preocupação ambiental e passa a integrar o conjunto de critérios relevantes para a tomada de decisão multicritério em IA.

2.11 Custo e Energia em Modelos de Linguagem

A avaliação de modelos de linguagem em contextos aplicados requer a consideração conjunta de desempenho e custo. Em especial, o custo computacional, a latência de resposta e o consumo energético exercem influência direta sobre a viabilidade de implantação, a

escalabilidade e a sustentabilidade das soluções baseadas em NLP (SCHWARTZ et al., 2020; ARGERICH; PATIÑO-MARTÍNEZ, 2024).

Em modelos de grande porte, o custo de inferência pode se tornar um fator limitante, sobretudo quando a aplicação exige múltiplas consultas, interações contínuas ou disponibilidade em tempo real. Nesses casos, o volume de parâmetros, o tamanho das entradas, a complexidade da arquitetura e a infraestrutura empregada afetam o tempo de resposta e a energia consumida por consulta (ARGERICH; PATIÑO-MARTÍNEZ, 2024; FERNANDEZ et al., 2025b). Assim, a adoção de um modelo não depende apenas da sua capacidade de produzir respostas de alta qualidade, mas também da eficiência com que essas respostas são geradas.

A literatura recente tem mostrado que a mensuração do custo energético em inferência é um tema central para a implementação responsável de sistemas de IA. Em particular, estudos voltados à inferência de LLMs evidenciam que o consumo pode variar significativamente conforme a configuração de hardware, a quantização do modelo, o paralelismo empregado e o tamanho da sequência processada. Dessa forma, avaliações comparativas entre modelos precisam incorporar condições operacionais realistas, evitando conclusões baseadas exclusivamente em cenários idealizados ou em métricas isoladas (ARGERICH; PATIÑO-MARTÍNEZ, 2024; FERNANDEZ et al., 2025b).

Quando observados em perspectiva educacional, esses fatores assumem importância adicional. Em Ambientes Virtuais de Aprendizagem, a adoção de modelos de linguagem tende a ocorrer em larga escala e com elevada frequência de uso, o que intensifica os custos totais associados à operação do sistema. Além disso, em contextos com limitações de infraestrutura, a exigência energética pode restringir a adoção de soluções mais robustas, tornando modelos menores ou mais otimizados alternativas mais adequadas. Nesse sentido, a eficiência computacional passa a ser um componente determinante da acessibilidade tecnológica (SCHWARTZ et al., 2020; TSCHAND et al., 2025).

Portanto, ao tratar de custo e energia em PLN e LLMs, observa-se que a escolha do modelo mais adequado deve considerar a relação entre desempenho e recursos consumidos. Essa perspectiva é especialmente importante quando o objetivo não é apenas maximizar resultados em uma tarefa isolada, mas selecionar uma solução sustentável, escalável e compatível com as restrições do ambiente de aplicação. Em consequência, custo e energia constituem dimensões centrais para a avaliação multicritério de modelos de linguagem em contextos educacionais.

2.12 Ajuste Fino

O ajuste fino é um procedimento pelo qual um modelo previamente treinado passa por uma etapa adicional de treinamento em um conjunto de dados específico, com o objetivo

de adaptá-lo a tarefas ou domínios particulares. Em termos gerais, essa abordagem permite que o modelo especialize seus parâmetros para responder de maneira mais adequada às demandas da tarefa-alvo, preservando, ao mesmo tempo, parte do conhecimento adquirido durante o pré-treinamento (ANISUZZAMAN et al., 2025; WU et al., 2025).

Entre as principais estratégias de ajuste fino supervisionado, destacam-se principalmente (ANISUZZAMAN et al., 2025):

- **Aprendizagem por transferência**: consiste em utilizar um modelo previamente treinado em grandes volumes de dados como base para uma nova tarefa específica. Dessa forma, o conhecimento já aprendido pelo modelo é reaproveitado, reduzindo a necessidade de grandes quantidades de dados rotulados e acelerando o treinamento.
- **Aprendizado multitarefa**: nessa abordagem, o modelo é treinado simultaneamente para diferentes tarefas relacionadas, compartilhando representações e padrões entre elas. Isso permite melhorar o desempenho geral e aumentar a capacidade de generalização do modelo.
- **Ajuste de instruções**: refere-se ao refinamento de modelos pré-treinados para que consigam compreender e seguir comandos explícitos de forma mais eficiente. Essa técnica é amplamente utilizada em tarefas como tradução automática, resumo de textos e resposta a perguntas.

Além do ajuste fino supervisionado, abordagens baseadas em aprendizado por reforço a partir de feedback humano (RLHF) também são utilizadas, nas quais o comportamento do modelo é refinado com apoio de avaliadores humanos, como modelagem de recompensa, a otimização de políticas proximal (PPO), a classificação comparativa e a realimentação de preferência. Em conjunto, essas técnicas procuram orientar o modelo para produzir saídas mais alinhadas à qualidade esperada, especialmente em aplicações práticas (ANISUZZAMAN et al., 2025).

Em um ajuste fino, várias etapas são envolvidas na preparação dos dados, incluindo a seleção do modelo pré-treinado mais apropriado e configuração cuidadosa dos hiperparâmetros. A adequação do conjunto de dados à tarefa, a escolha da arquitetura e o controle de parâmetros como taxa de aprendizado, tamanho do lote e número de épocas são aspectos essenciais para evitar sobreajuste e melhorar a generalização (ANISUZZAMAN et al., 2025; WU et al., 2025).

Assim, o ajuste fino se apresenta como uma estratégia relevante para especializar modelos de linguagem em aplicações específicas, oferecendo ganhos em conhecimento de domínio, desempenho, eficiência de dados e economia de recursos computacionais (ANISUZZAMAN et al., 2025; WU et al., 2025).

2.13 Engenharia de Prompt

A engenharia de prompt pode ser compreendida como o conjunto de práticas relacionadas à formulação, organização, otimização de instruções e configurações fornecidas a modelos de linguagem generativos, com o objetivo de direcionar suas respostas de maneira específica, eficiente e contextualizada. Nesse sentido, toda essa tarefa trabalha como uma forma de programação em linguagem natural, permitindo adaptar a interação com o modelo para obter saídas mais precisas, concretas, replicáveis e factualmente corretas (NACIMENTO, 2024; WHITE et al., 2023).

Os prompts podem variar em complexidade, desde solicitações simples até instruções mais elaboradas, no qual podem incluir exemplos e requisições de subdivisões da tarefa, o que auxilia o modelo na demanda. Além disso, os prompts podem ser classificados conforme sua estrutura, sendo eles instrutivos, de sistema, pergunta-resposta ou mistos, dependendo da estratégia empregada na formulação da entrada e na expectativa de resultado (NACIMENTO, 2024).

Apesar de sua relevância, prompts mal construídos podem comprometer a qualidade das respostas geradas, produzindo saídas vagas, ambíguas, tendenciosas ou incorretas, além da possibilidade de alucinação, termo empregado para a ação de quando há a geração de informações falsas ou imprecisas pelo modelo (NACIMENTO, 2024).

Outro processo de Engenharia de Prompt que vai além do texto de entrada é a configuração de execução dos modelos. Existem 4 principais parâmetros de configuração de geração de texto que influenciam diretamente o estilo, a coerência e a diversidade das saídas, sendo eles a Temperatura, *Top-k*, *Top-p* e *Min-p*.

A temperatura regula o grau de aleatoriedade na escolha de tokens durante decodificação, os valores baixos tendem a gerar saídas mais determinísticas e conservadoras, enquanto os valores altos, aumentam a criatividade, mas podem comprometer a coerência semântica (SACHETE et al., 2025). A Tabela 1 apresenta os valores de temperatura e quais os efeitos de cada valor na geração de texto.

Tabela 1 – Valores de temperatura e seus efeitos na geração de texto.

Valor da temperatura	**Comportamento esperado**
0,0	Geração totalmente determinística.
0,5	Saída moderadamente criativa, com relativa preservação da coerência.
1,0	Maior diversidade, mantendo coerência contextual.
1,5	Alta criatividade, com interpretações mais variadas.
2,0	Elevada criatividade, com maior risco de perda de coerência.

O parâmetro *top-p*, também conhecido como *nucleus sampling*, define o menor

conjunto de tokens cujas probabilidades acumuladas ultrapassam um valor limite P. Dessa forma, a seleção é adaptativa: em distribuições de probabilidade muito concentradas, poucos tokens serão considerados; em distribuições mais "achatadas", mais tokens entram no conjunto (HOLTZMAN et al., 2020). A Tabela 2 apresenta os valores de *top-p* e quais os efeitos de cada valor na geração de texto.

Tabela 2 – Valores de *top-p* e seus efeitos na geração de texto.

Valor de *top-p*	**Comportamento esperado**
0,3	Considera apenas tokens extremamente prováveis, produzindo respostas altamente conservadoras.
0,5	Seleciona somente os tokens mais prováveis, mantendo respostas ainda conservadoras.
0,7	Oferece um bom equilíbrio entre segurança e naturalidade na geração.
0,9	Permite diversidade moderada, preservando a coerência contextual.
1,0	Remove praticamente todas as restrições, aproximando-se da amostragem total.

O parâmetro *top-k* restringe a seleção de tokens aos k mais prováveis em cada etapa de geração. Assim, um valor menor de k torna a saída mais determinística, enquanto valores maiores permitem maior diversidade (FAN; LEWIS; DAUPHIN, 2018). A Tabela 3 apresenta os valores de *top-k* e quais os efeitos de cada valor na geração de texto.

Tabela 3 – Valores de *top-k* e seus efeitos na geração de texto.

Valor de *top-k*	**Comportamento esperado**
$k = 1$	Geração determinística, na qual apenas o token mais provável é escolhido.
$k = 10$	Produz respostas altamente controladas, com baixa diversidade.
$k = 50$	Proporciona diversidade moderada, sem comprometer significativamente a coerência.
$k = 100$	Aumenta a diversidade das respostas, com maior risco de inserção de termos fora de contexto.

O parâmetro *min-p* é uma técnica mais recente que define um limiar mínimo relativo à probabilidade do token mais provável. Através disso, um token só é incluído se sua probabilidade for maior que *p-min* $\times$ *p-max*. Esse mecanismo adapta o truncamento de acordo com a confiança do modelo, reduzindo tokens improváveis em cenários de alta incerteza (NGUYEN et al., 2025). A Tabela 4 apresenta os valores de *min-p* e quais os efeitos de cada valor na geração de texto.

De modo geral, a elaboração eficaz de prompts e configuração correta dos parâmetros é uma etapa muito importante, pois fornece instruções claras sobre o que o modelo deve seguir e tende a produzir respostas mais consistentes e alinhadas ao objetivo da tarefa.

Tabela 4 – Valores de *min-p* e seus efeitos na geração de texto.

Valor de *min-p*	**Comportamento esperado**
0,05	Configuração muito restritiva, considerando apenas tokens próximos ao mais provável.
0,10	Restrição moderada, com redução da probabilidade de incoerência.
0,20	Configuração mais flexível, preservando maior diversidade na geração.

2.14 Desafios e Limitações

Apesar dos avanços recentes e do crescente interesse no uso da Inteligência Artificial em Ambientes Virtuais de Aprendizagem (AVAs), a adoção de tais tecnologias ainda enfrenta uma série de desafios e limitações que precisam ser considerados de forma crítica. Esses obstáculos não se restringem apenas a aspectos técnicos, mas também abrangem dimensões éticas, pedagógicas, sociais e ambientais, as quais impactam diretamente a aceitação, a confiabilidade e a efetividade dessas ferramentas. Além disso, esses desafios se tornam ainda mais complexos quando a seleção da solução envolve múltiplos critérios, como desempenho, custo computacional, latência e consumo energético, exigindo decisões contextualizadas e metodologicamente justificadas.

Do ponto de vista ético, a rápida evolução da IA, em especial dos agentes conversacionais baseados em modelos de linguagem, demanda monitoramento contínuo para mitigar riscos associados a usos indevidos, enviesados ou antiéticos. Nesse sentido, torna-se essencial o estabelecimento de políticas institucionais, diretrizes pedagógicas e regulamentações em nível governamental que orientem o uso responsável da tecnologia, de modo a garantir transparência, equidade e proteção de dados sensíveis (ALI et al., 2024). A ausência dessas medidas pode favorecer a reprodução de vieses discriminatórios, a manipulação de informações e o uso inadequado dos sistemas em contextos educacionais.

No contexto educacional, um dos principais desafios está na natureza opaca dos modelos de linguagem, que frequentemente operam como "caixas-pretas" e dificultam a compreensão dos processos envolvidos em suas respostas e recomendações. Professores e estudantes muitas vezes não têm acesso aos critérios pelos quais determinado conteúdo foi selecionado ou priorizado, o que pode gerar insegurança em relação à confiabilidade dos resultados. Essa limitação se agrava diante do fenômeno conhecido como *alucinação*, em que o modelo produz informações plausíveis, mas incorretas ou irrelevantes, comprometendo a credibilidade da ferramenta e podendo afetar negativamente o processo de ensino-aprendizagem (ALIER et al., 2025; OUYANG et al., 2024).

Outro ponto crítico refere-se à integridade acadêmica. O uso de assistentes virtuais pode facilitar comportamentos oportunistas, como a utilização indevida de IA para resolver tarefas ou avaliações sem esforço cognitivo genuíno. Essa prática levanta sérias

preocupações sobre ética educacional, transparência e justiça avaliativa. Além disso, evidencia a necessidade de compreender as percepções dos estudantes sobre integridade acadêmica, de modo a promover o uso responsável e consciente dessas tecnologias no ambiente de ensino (ALI et al., 2024).

A eficácia das respostas fornecidas pelos modelos de linguagem também depende, em grande medida, da qualidade dos *prompts* utilizados. Usuários com menor familiaridade tecnológica podem enfrentar dificuldades para formular perguntas claras e bem estruturadas, o que reduz o potencial pedagógico da ferramenta. Essa exigência de interação mais sofisticada cria uma barreira de usabilidade que precisa ser superada por meio de estratégias de capacitação e de interfaces mais intuitivas (ALIER et al., 2025). Assim, embora esses sistemas possam apoiar a aprendizagem, sua efetividade permanece condicionada à capacidade do usuário de interagir adequadamente com o modelo.

Além disso, há preocupações crescentes quanto à possibilidade de a IA facilitar práticas de trapaça, plágio e reprodução acrítica de conteúdos. Tais práticas não apenas comprometem os princípios de equidade e inclusão, mas também podem reforçar desigualdades educacionais já existentes, sobretudo em contextos em que o acesso e o domínio da tecnologia não são uniformes (ALI et al., 2024). Assim, a dependência excessiva dessas ferramentas pode limitar a autonomia intelectual dos estudantes, prejudicando o desenvolvimento de competências fundamentais, como a criatividade, a reflexão crítica e a capacidade de resolução de problemas de forma independente (ALIER et al., 2025).

É importante destacar que, embora os assistentes virtuais baseados em IA possam desempenhar papéis semelhantes aos de tutores ou professores ao fornecerem explicações e orientações, ainda carecem da capacidade de compreender plenamente as necessidades individuais de aprendizagem e, sobretudo, os aspectos emocionais e sociais envolvidos no processo educativo. A ausência de interações humanas autênticas, essenciais para a motivação, a empatia e a construção de vínculos no ambiente escolar, representa uma limitação significativa e reforça o papel insubstituível do professor como mediador do aprendizado (ALIER et al., 2025; CHOI; KIM, 2024).

Em outro ponto, a eficácia de soluções baseadas em IA em AVAs é condicionada pela qualidade dos dados, pelos *benchmarks* utilizados e pelo contexto em que os modelos são avaliados. Em muitos casos, resultados obtidos em ambientes controlados não refletem integralmente a complexidade de uso em situações reais de ensino, nas quais variáveis como perfil do estudante, domínio de conteúdo, idioma, infraestrutura e objetivos pedagógicos interferem nos resultados. Dessa forma, a comparação entre modelos não deve ser reduzida a uma métrica isolada, pois diferentes configurações podem produzir desempenhos distintos conforme o cenário de aplicação, o que reforça a necessidade de abordagens multicritério para apoiar a decisão.

Por fim, emergem preocupações relacionadas à sustentabilidade ambiental associada

ao uso intensivo de modelos de IA. O crescimento da demanda por sistemas baseados em inteligência artificial implica maior consumo energético e maior pressão sobre recursos naturais, como energia elétrica e água utilizada em sistemas de resfriamento de data centers. Nesse contexto, torna-se fundamental desenvolver e adotar abordagens mais eficientes, incluindo modelos de menor escala e estratégias de otimização que reduzam o impacto ambiental dessas tecnologias (ECCARD; CAVALCANTE et al., 2025). Assim, a sustentabilidade passa a ser uma dimensão relevante não apenas do ponto de vista ecológico, mas também como parte da avaliação da viabilidade e da adequação dos modelos em cenários educacionais.

Dessa forma, os desafios e limitações associados ao uso de IA em AVAs evidenciam que a escolha da solução mais adequada não pode ser orientada por um único critério de desempenho. A complexidade do cenário educacional exige uma análise integrada de aspectos técnicos, éticos, pedagógicos e ambientais, o que reforça a relevância de um processo decisório estruturado, capaz de apoiar a seleção de modelos de forma transparente, contextualizada e sustentável.

3 Trabalhos Relacionados

Esta seção apresenta pesquisas relacionadas ao uso de modelos de linguagem em ambientes virtuais de aprendizagem, destacando seus objetivos, resultados e limitações, bem como a relevância desses trabalhos para a presente investigação.

Em sua pesquisa, Nutalapati, Velmurugan e Tiglao (2024) apresenta o *Coding Buddy*, um sistema de aprendizado on-line projetado para adaptar-se ao ritmo individual de cada estudante em ambientes virtuais de aprendizagem. O sistema monitora diferentes aspectos do comportamento dos alunos, como tempo gasto em cada aula, erros cometidos, número de tentativas, solicitações de ajuda e métricas refinadas de engajamento, incluindo movimentos do mouse e pressionamentos de tecla. Com base nesses dados, recomenda percursos de aprendizagem personalizados.

Para alcançar esse objetivo, o *Coding Buddy* integra um algoritmo de recomendação híbrido, que combina filtragem colaborativa e abordagens baseadas em conteúdo, além de um módulo de ajuste de dificuldade. O assistente de bate-papo do sistema é sustentado por um LLM (GPT-3.5), oferecendo apoio imediato ao aluno e fornecendo dados adicionais para o mecanismo adaptativo.

Os resultados foram expressivos: o grupo experimental obteve desempenho médio superior (82,3% contra 73,6% do grupo controle), maior taxa de conclusão do curso (91% contra 76%) e redução de 35% nos erros ao longo do tempo. Também se observou maior engajamento (7,2 horas semanais contra 5,1 do grupo controle) e maior retenção de conhecimento (76% após três meses, contra 61%). Apesar de os resultados gerais terem sido extremamente positivos, os autores destacam a necessidade de desenvolvimento de conteúdo mais sofisticado para alunos de alto desempenho, melhorias no assistente de IA e investigações de longo prazo.

Este estudo demonstra o potencial de modelos de linguagem em personalizar o aprendizado e otimizar a retenção, ainda que utilizando um LLM de grande porte, de custo elevado. Neste ponto, a presente pesquisa busca avançar ao avaliar SLMs, com a hipótese de manter desempenho relevante a um custo computacional reduzido.

Em sua pesquisa, Neumann et al. (2024) apresenta o desenvolvimento do *MoodleBot*, um *plugin* de código aberto integrado à plataforma Moodle. O objetivo foi oferecer feedback imediato sobre o conteúdo das aulas e tarefas, de modo a apoiar a aprendizagem autorregulada (*self-regulated learning* – SRL). Com um processo de conversação semelhante ao humano, o *MoodleBot* oferece aos alunos uma experiência próxima à interação com um tutor real, mas com as vantagens de respostas imediatas e disponibilidade 24 horas por dia, utilizando o modelo GPT-4 com engenharia de *prompts* inteligentes.

Os resultados indicaram boa aceitação: os alunos perceberam o *MoodleBot* como uma ferramenta útil para aprimorar sua jornada educacional, embora alguns estudantes ainda tenham preferido a mediação de um tutor humano. Em relação à precisão, 88% das respostas foram consideradas adequadas e corretas, porém a introdução de um mecanismo de verificação de fatos mais sofisticado poderia otimizar ainda mais essa taxa. Em termos de satisfação, 80% dos participantes afirmaram que utilizariam novamente a ferramenta, relatando que o assistente virtual contribuiu para uma experiência mais fluida e acessível. O custo por aluno variou entre $1,65 e $1,90 durante o período de utilização em uma palestra sobre tecnologia.

Os autores concluem que os chatbots podem mitigar a escassez de especialistas no ensino superior, mas ressaltam a necessidade de explorar outros modelos que possam oferecer vantagens de custo-benefício, mantendo alta precisão.

Esse trabalho relaciona-se diretamente à presente pesquisa, na medida em que demonstra a viabilidade de integrar chatbots a AVAs. Contudo, enquanto o *MoodleBot* explora um LLM de alto custo, esta pesquisa investiga SLMs, com foco em acessibilidade, eficiência e sustentabilidade computacional.

Outro estudo Choi e Kim (2025) aborda um ambiente de aprendizagem de programação sustentado por um LLM, projetado para oferecer suporte adaptativo e feedback personalizado aos códigos dos alunos, com diálogos e interações semelhantes às humanas. Os participantes podiam solicitar revisões de seus programas e esclarecimentos sobre dúvidas conceituais.

A pesquisa foi conduzida com abordagem mista, utilizando um desenho quase experimental fundamentado na teoria da autodeterminação. Os resultados apontaram que os alunos inseridos no ambiente baseado em LLM apresentaram melhorias significativas em sua capacidade de programação, especialmente aqueles com níveis moderados ou elevados de motivação autônoma. Também foi observada maior satisfação das necessidades psicológicas básicas em comparação com métodos tradicionais.

Entretanto, o estudo apresentou limitações, como o baixo número de participantes e o uso de medidas autorrelatadas para avaliar motivação e satisfação.

Essa pesquisa é relevante por destacar não apenas o desempenho técnico, mas também fatores motivacionais associados ao uso de modelos de linguagem. No entanto, assim como os demais trabalhos citados, baseia-se em LLMs de alta complexidade, o que contrasta com a proposta deste estudo de explorar SLMs em contextos educacionais, visando equilíbrio entre desempenho, custo e acessibilidade.

Pesquisas recentes têm investigado o potencial de Small Language Models (SLMs) em tarefas criativas, tradicionalmente dominadas por LLMs. O estudo de Marco, Rello e Gonzalo (2024) avaliou o modelo BART-large na geração de sinopses de filmes a partir de títulos, comparando seus resultados com produções humanas em cinco dimensões:

legibilidade, compreensibilidade, relevância, atratividade e criatividade. Com mais de 24 mil avaliações manuais de 68 participantes, os resultados mostraram que o SLM superou os textos humanos em todas as dimensões, exceto na criatividade. Além disso, os autores analisaram o impacto do viés dos avaliadores, revelando que o conhecimento sobre a autoria (humana ou IA) afetava negativamente a percepção da qualidade dos textos gerados por modelos. Por fim, comparações qualitativas com LLMs, como o GPT-4, indicaram que modelos maiores tendem a produzir textos mais consistentes, porém mais previsíveis, enquanto modelos menores, como o BART-large, geraram conteúdos mais originais e surpreendentes.

Esse trabalho é particularmente relevante para a presente pesquisa, pois demonstra que SLMs podem não apenas ser competitivos, mas também superar modelos maiores em determinadas dimensões de avaliação. Assim, reforça a hipótese de que modelos menores podem oferecer vantagens qualitativas e não apenas econômicas em contextos educacionais específicos.

O estudo de Chen et al. (2025) explorou o uso de LLMs na educação em engenharia, avaliando o desempenho de modelos como GPT-3.5 Turbo, GPT-4o e LLaMA 3 70B na correção de exercícios de análise de circuitos em nível de graduação. Para isso, foi criado um conjunto de dados com soluções oficiais e respostas reais de estudantes, convertidas em LaTeX, a fim de contornar limitações de reconhecimento de imagem. A avaliação considerou cinco métricas: completude, método, resposta final, erro aritmético e unidades. Os resultados indicaram que GPT-4o e LLaMA 3 70B superaram o GPT-3.5 Turbo em todas as métricas, cada um apresentando vantagens específicas.

O trabalho destacou ainda as limitações atuais dos LLMs e propôs *benchmarks* capazes de apoiar o desenvolvimento de tutores personalizados e confiáveis para cursos de engenharia, com potencial de aplicação em outras disciplinas. A relevância deste estudo para a presente pesquisa reside no estabelecimento de metodologias de avaliação rigorosas, que podem ser adaptadas para examinar o desempenho de SLMs em contextos educacionais variados.

O estudo de Bernardo et al. (2020) e trabalhos subsequentes investigam o impacto ambiental e o desempenho computacional de diferentes arquiteturas utilizadas em algoritmos de Aprendizado de Máquina (AM), evidenciando o custo energético associado ao treinamento e à execução de modelos de IA. Resultados de outros trabalhos indicam que o treinamento de um modelo de NLP pode apresentar níveis de emissão comparáveis aos de múltiplos veículos ao longo de seu ciclo de vida, enquanto experimentos mais recentes demonstram que execuções prolongadas de redes neurais convolucionais (RNCs) podem representar consumo energético significativo mesmo em cenários controlados. A análise comparativa entre arquiteturas revela *trade-offs* importantes: GPUs oferecem menor tempo de execução, porém apresentam maior consumo energético e pior eficiência em termos de *Energy-Delay*

Product (EDP) e emissão de CO_2e; CPUs tendem a apresentar melhor equilíbrio entre consumo e desempenho; e arquiteturas baseadas em ARM destacam-se pela baixa emissão, ainda que com maior tempo de processamento. Além disso, os resultados reforçam a relevância da otimização de algoritmos para diferentes plataformas, evidenciando que escolhas arquiteturais impactam diretamente tanto o desempenho quanto a sustentabilidade dos sistemas.

Esse trabalho contribui para a presente pesquisa ao evidenciar que a avaliação de modelos de linguagem não deve se restringir apenas à qualidade das respostas, mas também considerar aspectos computacionais e ambientais. Nesse sentido, reforça-se a importância de investigar SLMs como alternativas potencialmente mais eficientes, capazes de reduzir custos energéticos e impactos ambientais sem comprometer significativamente o desempenho em tarefas educacionais.

No contexto da aplicação de métodos multicritério de decisão, Pereira, Basilio e Santos (2026) apresentam a *pyDecision*, uma biblioteca abrangente na linguagem Python que implementa 70 métodos de decisão multicritério, incluindo AHP, TOPSIS, PROMETHEE, ELECTRE e VIKOR. O trabalho se destaca por integrar o ChatGPT como suporte à interpretação dos resultados, permitindo que os tomadores de decisão discutam e comparem os resultados dos diferentes métodos multicritério de forma intuitiva e interativa. Os autores concluem que, embora os modelos de linguagem sejam ferramentas indiscutivelmente poderosas para auxiliar na interpretação de dados complexos, suas respostas exigem verificação rigorosa por especialistas do domínio para evitar análises enganosas.

Enquanto a pesquisa de Pereira, Basilio e Santos (2026) utiliza a inteligência artificial generativa como um instrumento para facilitar a compreensão de uma análise multicritério, o presente trabalho aborda uma perspectiva complementar: a aplicação da decisão multicritério para avaliar e selecionar o modelo de linguagem mais adequado aos critérios definidos. Dessa forma, o *framework* proposto nesta pesquisa se apropria da robustez teórica dos métodos MCDA para estruturar a escolha da tecnologia, considerando simultaneamente critérios potencialmente conflitantes, como acurácia, tempo de resposta e consumo energético, de acordo com o nível de complexidade exigido pelo cenário educacional.

Em síntese, os trabalhos relacionados evidenciam que modelos de linguagem podem melhorar significativamente o engajamento, a personalização e a retenção em ambientes de aprendizagem. Todavia, observa-se a predominância de pesquisas baseadas em LLMs, que, embora apresentem alto desempenho, impõem limitações relevantes quanto ao custo, escalabilidade, latência e consumo de recursos computacionais. Além dos aspectos operacionais, emergem também preocupações relacionadas à sustentabilidade, uma vez que o uso intensivo está associado a elevado consumo energético e maiores emissões de carbono.

Nesse contexto, a presente pesquisa se diferencia ao investigar SLMs em comparação com LLMs, buscando compreender seu potencial de oferecer benefícios educacionais semelhantes, ao mesmo tempo em que promovem maior eficiência computacional, redução de custos operacionais e menor impacto ambiental. Mais do que comparar as métricas dos modelos de forma isolada, esta pesquisa busca preencher uma lacuna na literatura ao propor um *framework* de decisão multicritério. Por meio dessa abordagem orientada à decisão, acurácia, latência, custo computacional e consumo de energia deixam de ser variáveis independentes e passam a compor um mecanismo integrado de seleção, permitindo aos educadores e desenvolvedores a escolha do modelo mais adequado e sustentável para as necessidades específicas de cada tarefa educacional.

4 Metodologia

Esta pesquisa adota uma abordagem experimental comparativa, inserida na lógica da *Design Science Research* (DSR), com o objetivo de propor e avaliar um *framework* de decisão multicritério para seleção entre *Small Language Models* (SLMs) e *Large Language Models* (LLMs) em contextos educacionais. O artefato proposto busca orientar a escolha do modelo mais adequado a partir de evidências empíricas, considerando simultaneamente desempenho pedagógico, tempo de resposta, custo computacional e eficiência energética, de modo a explicitar os *trade-offs* entre qualidade e custo na tomada de decisão (KEENEY; RAIFFA, 1993; PEFFERS et al., 2007; HEVNER et al., 2008; BELTON; STEWART, 2012; HORITA; NETO; SANTOS, 2018; SCHWARTZ et al., 2020). A Figura 1 apresenta uma visão geral do DSR adotado nesta pesquisa.

A premissa central desta pesquisa é que não existe um modelo universalmente superior em todos os cenários. A adequação de um SLM ou de um LLM depende do contexto de uso, da disciplina avaliada, do nível de dificuldade da tarefa e das restrições operacionais do ambiente educacional. Nesse sentido, o *framework* é concebido como um artefato orientado à decisão, cuja função é transformar resultados experimentais em recomendações contextualizadas, reprodutíveis e justificáveis do ponto de vista científico.

4.1 Delineamento da pesquisa e lógica do artefato

Do ponto de vista metodológico, a pesquisa articula duas dimensões complementares. A primeira é a dimensão experimental, na qual diferentes modelos de linguagem são executados sobre conjuntos de questões educacionais em cenários controlados. A segunda é a dimensão de projeto do artefato, na qual os resultados obtidos alimentam uma função de utilidade multicritério capaz de apoiar a seleção do modelo mais adequado para cada cenário.

O processo de pesquisa foi estruturado em quatro etapas principais: (i) definição do artefato e dos critérios de decisão; (ii) construção das bases de dados de treinamento e avaliação; (iii) execução experimental dos modelos e coleta das métricas; e (iv) análise estatística e demonstração do *framework*. Essa organização é coerente com a lógica da DSR, na qual o artefato é concebido, instanciado, avaliado e refinado a partir de evidências empíricas (HEVNER et al., 2008; PEFFERS et al., 2007).

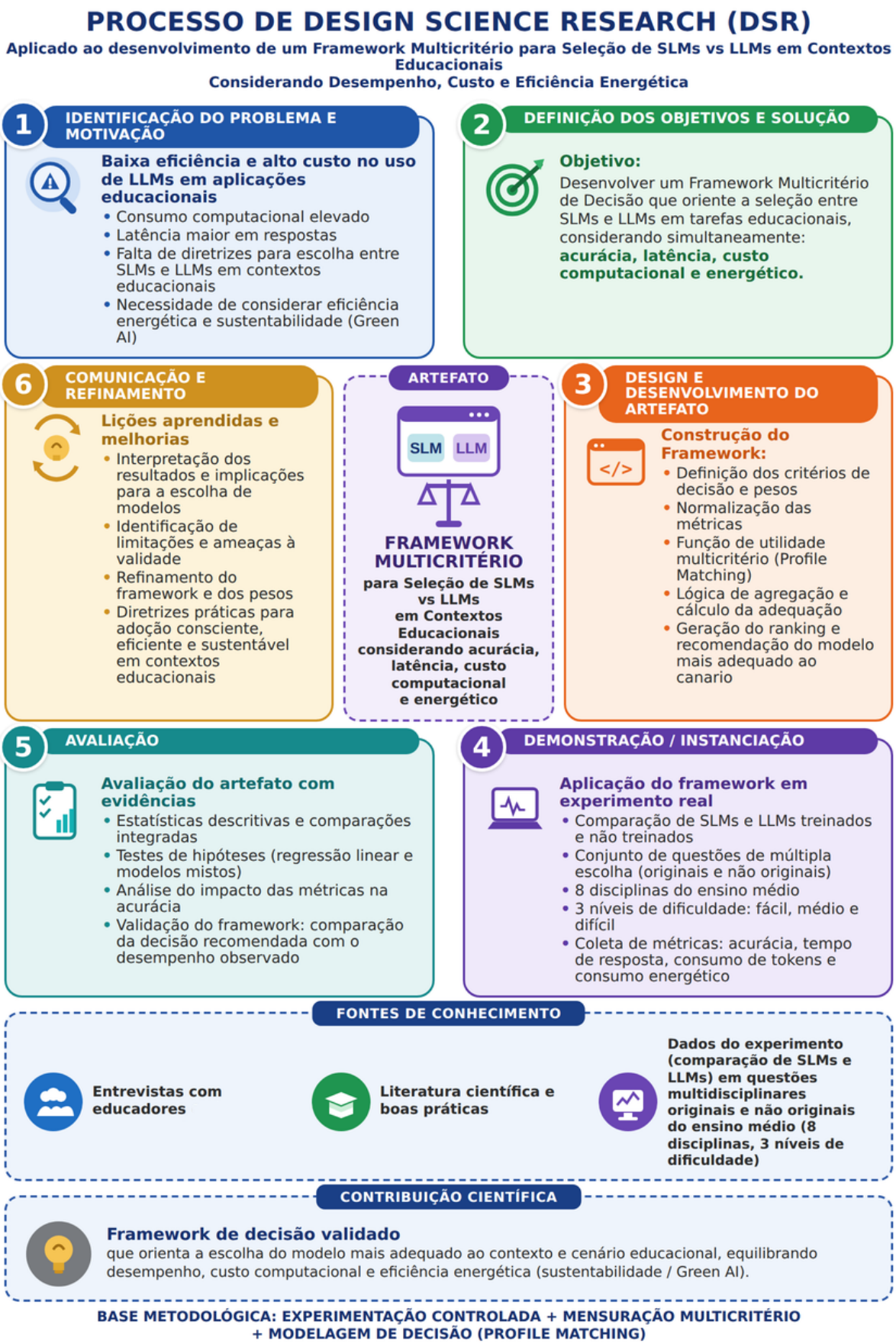


Figura 1 – *Design Science Research* da pesquisa.

4.2 Estrutura Experimental e Demonstração do *Framework*

O experimento de demonstração do *framework* foi conduzido em etapas sequenciais, descritas a seguir.

1. *Seleção dos modelos*: foram escolhidos modelos representativos das classes SLM e LLM, contemplando diferentes escalas de parâmetros e priorizando, sempre que possível, versões de código aberto ou de acesso reproduzível. A seleção buscou garantir diversidade suficiente para testar a sensibilidade do *framework* a variações de porte e capacidade dos modelos.

2. *Construção das bases de questões*: foram organizados dois conjuntos de avaliação. O primeiro é composto por questões não originais extraídas de exames oficiais do ENEM, enquanto o segundo é formado por questões originais elaboradas por especialistas. Em ambos os casos, as questões foram distribuídas por disciplina e por nível de dificuldade, permitindo avaliar o comportamento dos modelos em diferentes domínios cognitivos e contextos pedagógicos.

3. *Ajuste fino dos modelos*: os modelos SLMs selecionados foram submetidos a ajuste fino supervisionado, com o objetivo de verificar se o treinamento orientado por dados educacionais altera o desempenho e os custos computacionais do modelo. Essa etapa foi conduzida em ambiente de nuvem, com uso de bibliotecas apropriadas para ajuste fino eficiente de modelos quantizados.

4. *Configuração dos modelos para inferência*: os modelos foram padronizados quanto à forma de execução, à engenharia de *prompts* e aos parâmetros de geração, de modo a reduzir variações indevidas entre as interações e permitir comparabilidade entre as saídas.

5. *Execução dos modelos*: cada modelo respondeu ao mesmo conjunto de questões, com execução controlada das interações e randomização da ordem de apresentação, a fim de mitigar efeitos de posição e contexto. Durante a execução, foram coletadas métricas de acurácia, tempo de resposta, consumo de tokens e consumo energético.

6. *Armazenamento das saídas*: todas as respostas e metadados de execução foram registrados em banco de dados estruturado, incluindo o texto da resposta, a disciplina, o nível de dificuldade, a originalidade da questão, o modelo utilizado, o tempo de processamento, a quantidade de tokens e os indicadores de consumo energético.

7. *Construção do framework*: com base nas métricas coletadas, foram definidos os critérios de decisão, os pesos associados a cada critério e a lógica de agregação multicritério que compõe o *framework* proposto.

8. *Aplicação da função de decisão*: os resultados empíricos foram normalizados e inseridos na função de utilidade do *framework*, permitindo a comparação objetiva entre modelos e a obtenção de recomendações contextualizadas por cenário.

9. *Avaliação do artefato*: por fim, as decisões produzidas pelo *framework* foram comparadas com o desempenho observado experimentalmente, de modo a verificar se o modelo selecionado pela função de utilidade era de fato o mais adequado para cada cenário considerado.

A Figura 2 apresenta uma visão geral das etapas de estrutura experimental e demonstração do *framework*.

4.3 Bases de dados e cenários de demonstração

A pesquisa utiliza duas bases de dados principais: uma base para treinamento, utilizada no ajuste fino dos SLMs, e uma base para avaliação, empregada na demonstração do *framework*. Essa separação é essencial para evitar contaminação entre treinamento e teste e para garantir a integridade da comparação experimental.

A base de treinamento foi construída a partir da junção de três conjuntos públicos amplamente utilizados em tarefas educacionais e de raciocínio: **ScienceQA**, conjunto de dados coletado a partir dos currículos de ciências do ensino fundamental e médio, e contém 21.208 conteúdos multimodais de múltipla escolha envolvendo áreas de Biologia, Física, Geografia, História e derivados (LU et al., 2022); **RACE**, um conjunto de dados coletado a partir dos exames de inglês para alunos chineses do ensino fundamental e médio na faixa etária entre 12 e 18 anos, com perguntas geradas por especialistas humanos e aborda uma variedade de tópicos elaborados para avaliar a capacidade dos alunos de entender e raciocinar (LAI et al., 2017); e **GSM8K**, conjunto de dados focado em problemas de matemática de alta qualidade para o ensino fundamental, que demandam de duas a oito etapas para serem resolvidos e envolvem, principalmente, a realização de uma sequência de cálculos utilizando operações aritméticas básicas (COBBE et al., 2021).

Esses conjuntos, que possuem questões majoritariamente em inglês e multimodais, foram utilizados para adaptar os SLMs a tarefas de compreensão, raciocínio e resolução de problemas. A seleção de amostras respeitou a diversidade de domínios e a natureza educacional das questões, com o objetivo de produzir um ajuste fino compatível com o contexto da pesquisa.

A base de avaliação foi composta por questões de múltipla escolha distribuídas em oito disciplinas: História, Geografia, Matemática, Português, Inglês, Biologia, Física e Química. Para cada disciplina, foram selecionadas questões em três níveis de dificuldade: fácil, médio e difícil. Além disso, a avaliação foi organizada em duas fases:



Figura 2 – Etapas de estrutura experimental e demonstração do *framework*.

- questões não originais, extraídas de questões de bases públicas amplamente reconhecidas no contexto educacional brasileiro. As fontes selecionadas incluem provas oficiais do Exame Nacional do Ensino Médio (ENEM), considerando os anos de 2022, 2023 e 2024. A escolha por essas bases justifica-se pelo fato de representarem instrumentos consolidados de avaliação, elaborados com critérios pedagógicos rigorosos e alinhados às competências curriculares exigidas no ensino médio.
- questões originais, elaboradas por especialistas, apresentando questões originais que não foram utilizadas para o treinamento dos modelos, seguindo o mesmo rigor pedagógico.

Em ambas as fases, a distribuição contemplou questões de múltipla escolha com cinco alternativas de escolha, de forma a garantir a diversidade de formatos de avaliação. Ao todo, a base resultante conta com 240 questões (120 questões originais e 120 questões não originais) aplicadas uniformemente aos modelos.

4.4 Métricas

As métricas adotadas nesta pesquisa foram organizadas em três grupos: métricas de benefício, métricas de custo e métricas de avaliação.

4.4.1 Métricas de Benefício

As métricas de benefício representam os aspectos desejáveis do comportamento dos modelos, isto é, aqueles associados à qualidade da resposta gerada. Neste estudo, a principal métrica de benefício é a **corretude objetiva**.

A corretude objetiva, ou acurácia (A), do modelo m foi definida como a proporção de respostas corretas em relação ao total de questões avaliadas, conforme a Equação 4.1.

$$A_m = \frac{N_m^{\text{corretas}}}{N^{\text{total}}} \tag{4.1}$$

4.4.2 Métricas de Custo

As métricas de custo refletem o esforço computacional e energético necessário para que o modelo produza uma resposta. Neste trabalho, foram consideradas as métricas **tempo de resposta**, **custo computacional** e **custo energético**.

O tempo de resposta, ou latência (L), corresponde ao intervalo entre o envio da requisição ao modelo e o recebimento completo da resposta medidos em segundos, conforme a Equação 4.2.

$$L_m = t_m^{\text{recebimento}} - t_m^{\text{envio}} \tag{4.2}$$

O custo computacional (C) foi definido como a quantidade total de tokens consumidos durante a interação com o modelo, conforme a Equação 4.3.

$$C_m = T_m \tag{4.3}$$

O consumo energético (E) corresponde à energia total despendida durante o processo de inferência do modelo medidos em joules (J), conforme a Equação 4.4.

$$E_m = E_m^{CPU} + E_m^{GPU} \tag{4.4}$$

Essas métricas são fundamentais para a análise do *trade-off* entre desempenho e custo, especialmente em cenários de uso em larga escala.

4.4.3 Métricas de Avaliação

Como as métricas utilizadas possuem escalas distintas, foi necessária a definição de um procedimento de normalização para permitir sua agregação em uma única função de utilidade. De acordo com a Equação 4.5, a normalização adotada foi a *min-max*, que transforma os valores de cada métrica para o intervalo $[0, 1]$, em que x_m representa o valor da métrica para o modelo m, e $x_{\min}$ e $x_{\max}$ correspondem, respectivamente, aos valores mínimo e máximo observados entre todos os modelos avaliados.

$$\hat{x}_m = \frac{x_m - x_{\min}}{x_{\max} - x_{\min}} \tag{4.5}$$

Para métricas de benefício, como acurácia, o valor normalizado é interpretado como um limiar mínimo de suficiência operacional, e não como um alvo pontual a ser igualado exatamente. Dessa forma, valores acima do patamar definido são considerados plenamente satisfatórios, enquanto a penalização ocorre apenas quando o desempenho fica abaixo do nível desejado. Para métricas de custo, como tempo, tokens e energia, adota-se a interpretação inversa, de modo que menores valores representam melhor desempenho. Esse procedimento é compatível com abordagens multicritério e permite a comparação entre dimensões heterogêneas (KEENEY; RAIFFA, 1993; BELTON; STEWART, 2012; SAATY et al., 1980).

4.5 Função de Utilidade e Regras de Decisão

Com o objetivo de operacionalizar o *framework* proposto, esta pesquisa modela o problema de seleção entre *Small Language Models* (SLMs) e *Large Language Models* (LLMs) como um problema de decisão multicritério orientado a perfis de uso (*Profile Matching*). Em vez de adotar pesos predefinidos rígidos baseados em categorias abstratas, o artefato permite que a seleção seja balizada pelas restrições e metas reais do ambiente de implantação, focando exclusivamente nas quatro métricas centrais do estudo: Acurácia, Latência, Custo Computacional e Consumo Energético.

A premissa metodológica dessa modelagem é que a escolha do modelo ideal deve funcionar como uma combinação de requisitos (cenários). O decisor define o perfil desejado para a aplicação, e o *framework* atua como um mecanismo de otimização, retornando o modelo empírico que apresente a maior proximidade possível em relação aos parâmetros escolhidos, explicitando o *trade-off* de forma transparente e reprodutível.

4.5.1 Definição das Faixas de Utilização

Para viabilizar a combinação de escolhas, os resultados experimentais de todas as métricas foram organizados em faixas de utilização. Essa abordagem permite uma interpretação empírica consistente dos resultados, ao sintetizar a distribuição observada em intervalos qualitativos de uso. As faixas de utilização foram definidas em três níveis (Baixo, Médio e Alto) para cada critério, conforme detalhado a seguir:

- **Acurácia**
  - *Baixa (Até 40,00%):* Agrupa os modelos que falharam em compreender ou resolver as questões de forma satisfatória. Esta categoria captura, principalmente, modelos com menor capacidade de processamento ou versões que sofreram perda significativa de conhecimento após passarem pelo processo de treinamento.
  - *Média (De 40,01% a 65,00%):* Engloba o desempenho intermediário. Caracteriza-se como o patamar padrão alcançado pela maioria dos modelos de porte médio que não passaram por otimizações específicas para o contexto avaliado.
  - *Alta (Acima de 65,00%):* Isola o grupo com melhor desempenho no estudo. Captura os modelos que demonstraram real proficiência e capacidade de raciocínio nas tarefas educacionais, abrangendo tanto arquiteturas mais robustas quanto modelos menores excepcionalmente eficientes.
- **Latência:**

  - *Baixa / Rápida (Até 2,00s):* Representa a fluidez ideal para aplicações em tempo real. Engloba modelos altamente otimizados que conseguem processar e gerar respostas de forma quase instantânea.
  - *Média / Moderada (De 2,01s a 8,00s):* Indica um tempo de resposta aceitável para interações que toleram pequenos atrasos ou para processamentos assíncronos. Captura modelos maiores que mantêm boa eficiência, bem como modelos que sofreram leve perda de velocidade após serem ajustados.
  - *Alta / Lenta (Acima de 8,00s):* Representa um claro gargalo computacional. Engloba modelos excessivamente pesados ou arquiteturas que apresentaram anomalias e lentidão severa devido a um sobreajuste durante a fase de treinamento.

- **Custo Computacional:**
  - *Baixo (Até 450 tokens):* Evidencia as arquiteturas com alta capacidade de síntese, que entregam respostas diretas, objetivas e econômicas em termos de processamento.
  - *Médio (De 451 a 650 tokens):* Representa o volume de texto padrão gerado pela maioria dos modelos em seu estado original, que respondem de forma adequada sem serem excessivamente curtos ou prolongados.
  - *Alto (Acima de 650 tokens):* Isola o efeito de prolixidade, ou seja, respostas desnecessariamente longas. Nesta faixa, enquadram-se modelos que, seja por sua natureza ou por efeitos colaterais do treinamento, passaram a gerar um volume inflado e indesejado de texto.

- **Consumo Energético:**
  - *Baixo (Até 400 J):* Define as opções mais eficientes do ponto de vista energético e sustentável. São modelos viáveis para execução em equipamentos com limitações de infraestrutura e que se alinham aos princípios de *Green AI*.
  - *Médio (De 401 J a 1.500 J):* Representa uma faixa de transição. Agrupa modelos menores que passaram a consumir mais energia após o ajuste fino, além de modelos de maior porte que conseguem operar com uma eficiência energética surpreendente para o seu tamanho.
  - *Alto (Acima de 1.500 J):* Isola as arquiteturas que demandam recursos computacionais massivos. Inclui os modelos mais pesados do estudo e aqueles cujo processo de treinamento prejudicou severamente a sua eficiência na gestão de energia durante a geração de respostas.

Os valores-limite de cada faixa, para cada critério, foram definidos pelos autores com base em uma inspeção exploratória da distribuição dos dados obtidos na execução

dos modelos. Dessa forma, os intervalos adotados refletem padrões empíricos observados no experimento e foram estabelecidos de modo a preservar a interpretação ordinal entre os níveis baixo, médio e alto, sem aplicação de algoritmo formal de clusterização.

4.5.2 Modelagem Matemática de Aproximação de Perfil

A função de decisão do *framework* é realizada por meio do cálculo de distância multivariada entre as exigências do cenário (*target*) e os resultados empíricos dos modelos. Seja c um cenário educacional desejado, o decisor define um vetor de perfil alvo $P_c = (\tau_A, \tau_L, \tau_C, \tau_E)$, na qual cada τ_i representa o valor de referência normalizado da faixa (Baixa, Média ou Alta) de cada respectivo critério, sendo τ_A acurácia, τ_L latência, τ_C custo computacional e τ_E custo energético.

Para tornar a etapa de seleção reprodutível, cada categoria qualitativa foi codificada por um valor escalar fixo no intervalo normalizado $[0, 1]$. Assim, adotou-se o mapeamento $\tau(\text{Baixa}) = 0{,}00$, $\tau(\text{Média}) = 0{,}50$ e $\tau(\text{Alta}) = 1{,}00$, aplicado uniformemente a todos os critérios avaliados. Dessa forma, o perfil-alvo P_c passa a ser representado por um vetor numérico determinístico. Em paralelo, as métricas brutas extraídas experimentalmente de cada modelo m são normalizadas por meio da transformação *min-max* no intervalo $[0, 1]$ (conforme definido na Equação 4.5), gerando o vetor de desempenho $V_m = (\hat{A}_m, \hat{L}_m, \hat{C}_m, \hat{E}_m)$. A utilização da mesma escala normalizada para o perfil-alvo e para os resultados observados garante a comparabilidade direta entre os vetores utilizados no cálculo da distância.

Em seguida, o *framework* calcula a Distância Euclidiana $\Delta(m, c)$ entre o modelo analisado m e o perfil alvo c, conforme a Equação 4.6:

$$\Delta(m, c) = \sqrt{(\hat{A}_m - \tau_A)^2 + (\hat{L}_m - \tau_L)^2 + (\hat{C}_m - \tau_C)^2 + (\hat{E}_m - \tau_E)^2} \tag{4.6}$$

A escolha da distância euclidiana (norma L2) como função de agregação decorre de três considerações. Primeiro, a L2 é a métrica natural em problemas de aproximação de perfil e em métodos de ponto de referência da literatura de MCDA, aparecendo como núcleo do TOPSIS (BELTON; STEWART, 2012), no qual alternativas são avaliadas pela distância euclidiana até uma solução ideal. Segundo, a L2 penaliza desvios grandes de forma quadrática, o que é adequado ao contexto educacional: um modelo que falhe severamente em um critério (por exemplo, latência muito acima do tolerado em um cenário de tutoria em tempo real) deve ser mais penalizado do que um modelo que apresente pequenos desvios distribuídos entre vários critérios, comportamento que a norma L1 não captura. Terceiro, a L2 preserva a interpretabilidade geométrica do espaço multicritério normalizado em $[0, 1]$, permitindo que a distância seja lida diretamente como aderência global do modelo ao perfil-alvo.

A decisão multicritério consiste em solucionar o problema de minimização dessa distância. A Equação 4.7 representa a função de seleção do modelo ótimo m^*, ou seja, aquele que possui a maior aderência (menor distância Δ) às restrições do cenário escolhido.

$$m^*(c) = \arg\min_{m \in M} \Delta(m, c) \tag{4.7}$$

Para ilustrar a aplicação deste *framework*, a Tabela 5 apresenta exemplos de simulações de cenários de uso reais em ambientes educacionais desenvolvido pelos próprios autores, mapeando a combinação de exigências para os quais o algoritmo buscaria o modelo com a menor distância Δ.

Ressalta-se que o *framework* proposto não emprega pesos por critério no sentido clássico da MCDA compensatória. Os quatro critérios considerados (acurácia, latência, custo computacional e consumo energético) contribuem de forma equivalente para o cálculo da distância euclidiana. Assim, o elemento definido pelo decisor não são pesos numéricos, mas o *perfil-alvo* do cenário, isto é, o nível desejado (Baixo, Médio ou Alto) para cada critério. Essa modelagem não elimina a subjetividade inerente ao processo decisório, mas a desloca para uma escolha explícita, transparente, rastreável e passível de replicação. Em vez de ocultar preferências em um vetor implícito de pesos, o *framework* estrutura o problema como uma comparação direta entre perfis operacionais previamente definidos e o desempenho observado dos modelos, permitindo que a justificativa da decisão esteja associada aos requisitos do cenário educacional e proporcionando maior transparência e reprodutibilidade na análise multicritério.

4.6 Hipóteses e Variáveis

As hipóteses desta pesquisa foram estruturadas de modo a validar a eficácia de um *framework* de apoio à decisão multicritério, capaz de selecionar o modelo empírico mais adequado por meio da aproximação de perfis de uso (*Profile Matching*), não somente o desempenho isolado de modelos de linguagem.

As variáveis independentes do experimento incluem o tipo de modelo, o porte do modelo, o fato de o modelo ter sido treinado ou não, a disciplina, o nível de dificuldade e a originalidade da questão. As variáveis dependentes são acurácia, tempo de resposta, consumo de tokens e consumo energético.

4.6.1 Dimensão 1: Desempenho por Complexidade

- **H1**: SLMs de porte médio apresentam desempenho equivalente a LLMs em tarefas de nível Fácil e Médio, quando avaliados em termos de acurácia.

Tabela 5 – Cenários de uso e respectivas configurações de busca no *framework*

Cenário Educacional	**Acurácia**	**Latência**	**Custo (Tokens)**	**Energia**
Tutoria em Tempo Real *(Feedback instantâneo via chat)*	Média	Baixa	Baixo	Médio
Correção em Lote (*Batch*) *(Processamento assíncrono de atividades)*	Alta	Alta	Médio	Médio
Simulado ENEM com Feedback Explicativo *(Respostas detalhadas e maior confiabilidade)*	Alta	Média	Médio	Alto
Personalização Adaptativa de Trilhas *(Recomendação e adaptação contínua)*	Alta	Média	Médio	Médio
Sala de Aula Premium Multimodal *(Experiência rica e responsiva)*	Alta	Baixa	Alto	Alto
Quiz Gamificado Mobile *(Aplicação leve para celular)*	Baixa	Baixa	Baixo	Baixo
Escola com Recursos Limitados *(Conectividade instável e infraestrutura restrita)*	Média	Média	Baixo	Baixo
Monitoria Escalável para Grande Volume *(Muitos usuários simultâneos)*	Média	Baixa	Médio	Alto
Raciocínio Lógico Complexo *(Sem restrição de hardware)*	Alta	Média	Alto	Alto
Baseline Equilibrado para AVA Geral *(Cenário neutro para comparação)*	Média	Média	Médio	Médio

Nota. Cada cenário submetido ao *framework* representa um vetor de metas multicritério relacionado às métricas de acurácia, latência, custo computacional e consumo energético. O algoritmo percorre a base empírica do experimento e retorna o modelo com menor distância em relação ao perfil-alvo definido para aquele cenário.

- **H2**: LLMs apresentam desempenho superior em tarefas de nível Difícil, especialmente em cenários que exigem maior capacidade de raciocínio e generalização.

4.6.2 Dimensão 2: Eficiência Computacional

- **H3**: SLMs apresentam menor tempo de resposta em comparação aos LLMs.
- **H4**: SLMs apresentam menor consumo de tokens em comparação aos LLMs.
- **H5**: SLMs apresentam menor consumo energético em comparação aos LLMs.
- **H6**: O custo computacional aumenta de forma não linear com o aumento do porte do modelo.
- **H7**: O custo energético aumenta de forma não linear com o aumento do porte do modelo.

4.6.3 Dimensão 3: *Trade-offs* entre Desempenho e Custo

- **H8**: Existem *trade-offs* significativos entre acurácia e custo computacional, variando conforme disciplina, nível de dificuldade e originalidade da questão.
- **H9**: Modelos treinados apresentam ganho de desempenho em relação aos não treinados, embora esse ganho possa implicar variações no custo computacional e energético.

4.6.4 Dimensão 4: Adequação Contextual e Decisão

- **H10**: Não existe um modelo universalmente superior em todos os cenários operacionais; a escolha do modelo ideal depende da combinação específica de exigências de cada ambiente de aplicação.
- **H11**: A aplicação de um *framework* de decisão multicritério baseado em aproximação de perfis (*Profile Matching*) permite selecionar modelos mais adequados à realidade do usuário do que uma escolha baseada na maximização de métricas isoladas.

Essas hipóteses permitem avaliar não apenas o comportamento individual dos modelos frente às variáveis experimentais, mas também validar a capacidade do *framework* proposto em capturar e operacionalizar os *trade-offs* existentes, minimizando a distância entre as restrições reais de um cenário educacional e o desempenho empírico das arquiteturas.

4.7 Procedimentos experimentais

Os procedimentos experimentais foram organizados em três fases: preparação, intervenção e análise. Essa organização teve como objetivo garantir a padronização da infraestrutura experimental, a execução controlada dos experimentos e a aplicação de métodos analíticos consistentes para avaliação dos modelos de linguagem no contexto educacional.

A Figura 3 apresenta uma visão geral das etapas que compõem os procedimentos experimentais adotados nesta pesquisa.

4.7.1 Fase de Preparação

Antes da realização do experimento, conduziu-se uma fase de preparação da infraestrutura do *framework*, visando garantir reprodutibilidade e padronização do processo experimental.

4.7.1.1 Implementação da Aplicação

Na primeira etapa de preparação, foi desenvolvida uma infraestrutura para demonstração do *framework*.

A arquitetura deste sistema compreendeu um script em Node.js, ambiente de execução de código JavaScript no lado do servidor, que possui grande consistência, segurança, escalabilidade e desempenho.

Para o armazenamento dos dados, utilizou-se o banco de dados relacional MySQL, devido à sua alta disponibilidade, segurança, facilidade de uso e bom desempenho. Foram projetadas três tabelas principais: **Perguntas** (Tabela 6); **Respostas** (Tabela 7); e **Energia** (Tabela 8).

Tabela 6 – Estrutura da tabela Perguntas

Campo	**Descrição**
Texto da pergunta	Armazena o enunciado integral da pergunta.
Disciplina	Registra a área de conhecimento à qual a pergunta pertence.
Nível de dificuldade	Indica o grau de complexidade atribuído à pergunta (Fácil, Médio ou Difícil).
Alternativas de resposta	Contém as alternativas de resposta associadas à pergunta.
Pergunta original	Informa a originalidade da questão (Sim ou Não).

Esse armazenamento viabilizou a coleta dos dados necessários para demonstração do *framework*.

4.7 – PROCEDIMENTOS EXPERIMENTAIS

Os procedimentos experimentais foram organizados em três fases: preparação, intervenção e análise.

1 FASE DE PREPARAÇÃO
Preparação da infraestrutura do framework e dos recursos do experimento.

2 FASE DE INTERVENÇÃO
Execução controlada dos modelos e coleta de métricas.

3 FASE DE ANÁLISE
Tratamento dos dados, aplicação do framework e avaliação do artefato.

4.7.1 FASE DE PREPARAÇÃO

4.7.1.1 Implementação da Aplicação

Infraestrutura Node.js/MySQL criada para validação experimental do framework.

4.7.1.2 Modelos de linguagem

Seleção e classificação de SLMs/LLMs para comparação experimental.

4.7.1.3 Ajuste fino dos Modelos

Ajuste fino supervisionado aplicado aos SLMs com LoRA/QLoRA.

4.7.1.4 Engenharia de prompts e inferência

Padronização de prompts e parâmetros para inferência comparável.

Resultado da Preparação: Infraestrutura pronta, reprodutível e padronizada para os experimentos.

4.7.2 FASE DE INTERVENÇÃO

Execução dos Experimentos

- Execução dos modelos (SLMs e LLMs treinados e não treinados)
- Questões de múltipla escolha (originais e não originais)
- 8 disciplinas e 3 níveis de dificuldade (fácil, médio, difícil)

Métricas Coletadas

- Acurácia (taxa de acerto)
- Tempo de resposta
- Consumo de tokens
- Consumo energético (J)

Resultado da Intervenção: Métricas empíricas coletadas para todos os modelos e cenários.

4.7.3 FASE DE ANÁLISE

4.7.3.1 Estrutura dos dados experimentais

Organização dos dados experimentais coletados das execuções dos modelos educacionais.

4.7.3.2 Variáveis analisadas

Definição das variáveis dependentes e independentes avaliadas durante os experimentos.

4.7.3.3 Estratégia de análise estatística

Aplicação de métodos estatísticos robustos para análise multidimensional dos resultados.

4.7.3.4 Análise das hipóteses

Avaliação estatística das hipóteses relacionadas desempenho, custo e adequação contextual.

FLUXO GERAL DO EXPERIMENTO

Infraestrutura preparada → Bases de questões e modelos prontos → Execução controlada dos modelos → Coleta de métricas (4 dimensões) → Dados armazenados e tratados → Aplicação do framework multicritério → Resultado e validação do artefato

OBJETIVO DOS PROCEDIMENTOS EXPERIMENTAIS

Conduzir experimentos controlados para gerar evidências empíricas sobre o desempenho, custo computacional e eficiência energética de SLMs e LLMs, permitindo a validação de um framework multicritério para seleção do modelo mais adequado ao contexto educacional.

Figura 3 – Estrutura geral dos procedimentos experimentais da pesquisa.

Tabela 7 – Estrutura da tabela Respostas

Campo	**Descrição**
ID da pergunta	Identificador da pergunta associada à resposta registrada.
Resposta gerada pelo modelo	Armazena a resposta integral gerada pelo modelo.
Quantidade de tokens utilizados	Registra o número de tokens consumidos na geração da resposta.
Data e horário do envio	Informa o momento em que a requisição foi enviada ao modelo.
Data e horário do recebimento	Registra o momento em que a resposta foi recebida pelo sistema.
Modelo utilizado	Identifica o modelo utilizado na geração da resposta.
Modelo treinado	Indica se o modelo utilizado foi treinado (Sim ou Não).

Tabela 8 – Estrutura da tabela Energia

Campo	**Descrição**
Modelo utilizado	Identifica o modelo executado na medição de energia.
Modelo treinado	Indica se o modelo utilizado foi treinado (Sim ou Não).
Perguntas originais	Informa se o conjunto de entrada utilizou perguntas originais (Sim ou Não).
Energia CPU	Registra o consumo energético associado ao processamento em CPU em joules.
Energia GPU	Registra o consumo energético associado ao processamento em GPU em joules.
Energia total	Corresponde ao consumo energético agregado da execução em joules.
Watts médio CPU	Indica a potência média observada na CPU em watts.
Watts médio GPU	Indica a potência média observada na GPU em watts.
Watts médio total	Indica a potência média total observada durante a execução em watts.
Watts máximo CPU	Registra o maior valor de potência observado na CPU em watts.
Watts máximo GPU	Registra o maior valor de potência observado na GPU em watts.
Watts máximo total	Registra o maior valor de potência total observado durante a execução em Watts.

4.7.1.2 Modelos de linguagem

Na segunda etapa de preparação, foram selecionados os SLMs e LLMs para compor a demonstração do *framework* da pesquisa.

Os modelos foram classificados pelos autores de acordo com seu tamanho em bilhões de parâmetros:

- *Pequeno*: até 3 bilhões de parâmetros;
- *Médio*: entre 3 bilhões e 8 bilhões de parâmetros;
- *Grande*: acima de 8 bilhões de parâmetros.

Como a literatura atual não define de forma consensual o limite exato de número de parâmetros para um SLM ser categorizado, considerou-se, neste estudo, que modelos com até 8 bilhões de parâmetros seriam tratados como SLMs, enquanto modelos com mais de 8 bilhões de parâmetros seriam considerados LLMs. Essa classificação não pretende estabelecer um corte definitivo na literatura, mas apenas operacionalizar a comparação experimental entre categorias de diferente porte.

A seleção foi guiada por cinco critérios definidos pelos autores:

1. Modelos recentes e de relevância significativa na comunidade científica ou industrial;
2. Permissão para ajuste fino;
3. Acesso via API (preferencialmente RESTful);
4. Documentação acessível e facilidade de integração com Node.js.
5. Acesso *open-source*.

A Tabela 9 apresenta os modelos selecionados após o estágio de levantamento e análise, classificando-os por modelo, fabricante, número de parâmetros em bilhões, arquitetura e classificação.

A seguir, apresenta-se uma breve descrição de cada modelo, com justificativa de sua inclusão no estudo:

O **LLaMA 3.2** (Meta) contempla variantes compactas (por exemplo, 1B) voltadas a cenários de *edge* e *deploys* privados, sendo instrução-afinadas e multilíngues. Essas versões foram concebidas para facilitar a execução local ou em infraestruturas restritas, mantendo desempenho satisfatório em tarefas como diálogo, sumarização e garantia de qualidade — características relevantes para contextos educacionais com limitação de recursos (LLAMA, 2024). A utilização da variante de 1 bilhão de parâmetros permite investigar a viabilidade

Tabela 9 – Modelos de linguagem selecionados para experimentação

Modelo	**Fabricante**	**Parâmetros**	**Arquitetura**	**Classificação**
LLaMA 3.2 1B	Meta AI	1B	Transformer Decoder	Pequeno
StableLM 2 1.6B	Stability AI	1,6B	Transformer Decoder	Pequeno
Gemma 2 2B	Google DeepMind	2B	Transformer Decoder	Pequeno
Phi 3.5 mini 3.8B	Microsoft	3,8B	Transformer Decoder	Médio
Qwen 3 4B	Alibaba Cloud	4B	Transformer Decoder	Médio
Mistral 7B	Mistral AI	7B	Dense Transformer Decoder	Médio
Qwen 2.5 7B	Alibaba Cloud	7B	Transformer Decoder	Médio
LLaMA 3.1 8B	Meta AI	8B	Transformer Decoder	Médio
Gemma 3 12B	Google DeepMind	12B	Transformer Decoder	Grande
GPT OSS 20B	OpenAI	20B	Mixture-of-Experts (MoE)	Grande
Gemma 3 27B	Google DeepMind	27B	Transformer Decoder	Grande
Qwen 3 30B	Alibaba Cloud	30B	Transformer Decoder	Grande
LLaMA 3.1 70B	Meta AI	70B	Transformer Decoder	Grande

de modelos *open-source* muito compactos como alternativas práticas em ambientes escolares ou corporativos com restrições computacionais, além de prover um *baseline* reproduzível e amplamente reconhecido na literatura comparativa.

O **StableLM 2** (Stability AI) é um SLM com cerca de 1,6 bilhão de parâmetros, disponibilizado em versões base e instrução-afinadas, treinado sobre grandes volumes de tokens. Devido à sua licença permissiva e ao foco em eficiência, o StableLM 2 tornou-se uma referência relevante para estudos que avaliam *trade-offs* entre custo, velocidade de inferência e qualidade de resposta em contextos práticos (STABILITY, 2024). Sua inclusão é importante para comparar SLMs *open-source* treinados em larga escala de tokens frente a alternativas comerciais, permitindo analisar como diferenças em corpora e estratégias de ajuste fino impactam a correção e a clareza das explicações geradas.

O **Gemma 2 (2B)** representa uma renovação da linha de modelos desenvolvida pela Google DeepMind, focada em eficiência e execução em dispositivos com recursos limitados. Com cerca de 2 bilhões de parâmetros, esse modelo foi projetado para balancear desempenho e baixo custo de inferência, mantendo boa compreensão linguística e capacidade de geração coerente (JI; KUMAR, 2024). Sua inclusão no estudo permite comparar o progresso evolutivo entre gerações (Gemma 2 vs. Gemma 3), destacando como aprimoramentos arquiteturais e de treinamento afetam a precisão e o custo computacional em tarefas educacionais.

A família **Phi-3.5** inclui a variante *Phi-3.5-mini* (3,8 bilhões de parâmetros), caracterizada por contexto extenso e foco em inferência eficiente, com robustez em tarefas de raciocínio e explicação (TRUFINESCU, 2024). Ao integrar a Phi-3.5-mini ao grupo médio, busca-se analisar o ganho incremental de capacidade ao escalar de 2–3B para 4B parâmetros dentro de uma mesma linha arquitetural, avaliando como isso se traduz em correção, coerência das explicações e tempo de resposta em questões de diferentes complexidades. Essa comparação é particularmente relevante porque mantém continuidade

arquitetural, permitindo isolar efeitos atribuíveis ao aumento de parâmetros e à melhor gestão de contexto.

O **Mistral 7B** (Mistral AI) é um modelo denso de aproximadamente 7,3 bilhões de parâmetros, reconhecido por seu custo-benefício e desempenho competitivo em benchmarks padronizados. O projeto enfatiza eficiência arquitetural e otimizações de inferência, frequentemente alcançando resultados comparáveis a modelos maiores em tarefas de geração e compreensão (MISTRALAI, 2023). Sua presença no conjunto de comparação serve como caso representativo de design eficiente na faixa de 7B, permitindo avaliar se ganhos arquiteturais e de engenharia podem competir com aumentos de escala em termos de corretude e velocidade.

O **Qwen 2.5 (7B)** é um modelo *open-source* desenvolvido pela Alibaba Cloud, com cerca de 7 bilhões de parâmetros, treinado em múltiplos idiomas e ajustado para tarefas de raciocínio, compreensão e geração contextual (CLOUD, 2024). O modelo combina uma arquitetura otimizada e vocabulário amplo, resultando em desempenho competitivo em benchmarks gerais e específicos de QA. Sua inserção neste estudo visa comparar sua eficiência e coerência textual em relação a modelos de porte semelhante (como o Mistral 7B), explorando diferenças derivadas de treinamento multilíngue e abordagens de ajuste fino voltadas à precisão semântica.

O **LLaMA 3.1 (8B)** é uma variante intermediária da família LLaMA, projetada para oferecer equilíbrio entre capacidade de raciocínio e custo de inferência (IA, 2024). Com cerca de 8 bilhões de parâmetros, o modelo foi otimizado para instruções, chat e tarefas educacionais, apresentando desempenho robusto em diversas línguas e contextos. Sua inclusão neste estudo tem o propósito de avaliar o comportamento de um modelo de porte médio-alto da Meta frente a outras arquiteturas, investigando como melhorias progressivas nas versões da série LLaMA impactam o desempenho em tarefas de múltipla escolha e respostas discursivas.

A **família Gemma 3 (12B)** representa uma faixa intermediária de grande relevância, equilibrando custo computacional e desempenho em tarefas complexas (GOOGLE, 2025). A variante de 12 bilhões de parâmetros foi projetada para maximizar coerência e raciocínio sem sacrificar significativamente a eficiência, sendo adequada para aplicações de QA, ensino assistido e geração de conteúdo explicativo. Sua integração ao conjunto experimental permite observar o comportamento de modelos de médio porte dentro da mesma linha arquitetural, elucidando o impacto incremental do aumento de escala sobre corretude, clareza argumentativa e estabilidade das respostas.

O **GPT OSS 20B** é um modelo *open-source* de aproximadamente 20 bilhões de parâmetros, inspirado em arquiteturas semelhantes às da linha GPT, com foco em generalização e capacidade de instrução (OPENAI, 2025). Treinado sobre corpora extensos e diversificados, busca reproduzir o comportamento de LLMs proprietários em ambiente

aberto, mantendo transparência e possibilidade de reprodução. Sua inclusão no estudo visa oferecer um ponto de comparação entre soluções abertas de grande escala e alternativas comerciais, permitindo analisar em que medida o tamanho e a abertura do modelo impactam a qualidade das respostas e o custo energético de execução.

A **família Gemma 3 (27B)** representa a continuidade do desenvolvimento Gemma, com variantes de maior porte (por exemplo, 27 bilhões de parâmetros) projetadas para lidar com entradas longas, oferecer quantizações eficientes e manter bom desempenho multilíngue em tarefas de resumo, QA e raciocínio lógico (GOOGLE, 2025). A variante de 27B é útil para explorar o comportamento de modelos "grandes, porém acessíveis", que buscam democratizar a capacidade de modelos robustos sem escalar para centenas de bilhões de parâmetros. Sua inclusão permite comparar ganhos de escala em cenários educacionais que demandam compreensão profunda e coerência em respostas discursivas.

A **família Qwen 3** (Alibaba Cloud) inclui variantes de diferentes escalas, como os modelos de 4 bilhões e 30 bilhões de parâmetros, projetados para oferecer desempenho robusto em tarefas de raciocínio, compreensão contextual e geração de texto multilíngue (CLOUD, 2025). A variante de 4 bilhões de parâmetros apresenta-se como uma alternativa eficiente para cenários com restrições computacionais, buscando equilibrar qualidade de resposta e custo de inferência, enquanto a versão de 30B amplia significativamente a capacidade de raciocínio, coerência discursiva e processamento de contextos longos. A inclusão conjunta dos modelos Qwen 3-4B e Qwen 3-30B permite realizar uma análise comparativa dentro da mesma família arquitetural, investigando o impacto do aumento de escala sobre métricas como acurácia, consistência das respostas e custo computacional. Essa abordagem é particularmente relevante para isolar os efeitos do número de parâmetros e das otimizações internas do modelo.

O **LLaMA 3.1 (70B)** (Meta) é um LLM de grande porte (na ordem de dezenas de bilhões de parâmetros) otimizado para instruções e diálogo multilíngue, tendo se destacado em benchmarks por seu desempenho em tarefas de geração e compreensão contextual (IA, 2024). A disponibilidade de variantes oficiais e de repositórios de modelos *open-weight* torna o LLaMA 3.1 particularmente valioso para estudos comparativos que exigem reprodutibilidade e possibilidade de *deploy self-hosted* para mensuração de custos de inferência. Sua inclusão permite avaliar o impacto do aumento substancial de parâmetros sobre precisão, coerência explicativa e custos operacionais em comparação com modelos comerciais fechados.

Todos os modelos utilizados neste estudo foram obtidos em suas versões originais disponibilizadas pelos respectivos desenvolvedores por meio da plataforma Hugging Face (<https://huggingface.co>). Posteriormente, os modelos foram convertidos para o formato GGUF (*GPT-Generated Unified Format*), um padrão binário otimizado para execução local eficiente, amplamente utilizado em *frameworks* de inferência leve. Durante esse processo,

foi aplicada quantização do tipo Q8 (técnica de compressão que reduz o tamanho de Modelos de Linguagem), visando reduzir o consumo de memória e o custo computacional, ao mesmo tempo em que se preserva, de forma satisfatória, a qualidade das respostas geradas.

4.7.1.3 Ajuste fino dos Modelos

O treinamento dos modelos SLMs foi realizado com o objetivo de observar as diferenças de desempenho, acurácia, consumo energético e os principais *trade-offs*, sendo uma condição experimental para pesquisa, utilizando a técnica de ajuste fino supervisionado. A Figura 4 apresenta uma visão geral do processo empregado, incluindo a seleção dos modelos, preparação dos dados educacionais, ambiente de treinamento e os resultados esperados após o ajuste fino.

Como o ambiente local disponível não possuía poder de processamento necessário para realizar o ajuste fino desses modelos de forma viável, optou-se pelo uso do Google Colab, uma ferramenta com um ambiente Jupyter hospedado na nuvem que disponibiliza acesso a GPUs e TPUs, conforme a disponibilidade da plataforma. Durante todo o experimento, o treinamento foi executado em uma GPU T4, disponibilizada através do plano gratuito, adequada para cargas de trabalho de aprendizado profundo em cenários com recursos limitados (Google Research, 2024). Em todo o script de treinamento foi utilizada a linguagem Python.

Para a implementação do treinamento, foram utilizadas as bibliotecas `unsloth`, `datasets`, `transformers` e `trl`. A biblioteca `transformers` fornece a infraestrutura de treinamento baseada em *Trainer* e *TrainingArguments*, enquanto o *SFTTrainer*, da biblioteca TRL, é voltado especificamente para ajuste fino supervisionado em modelos de linguagem autoregressivos. O Unsloth foi empregado por oferecer uma implementação otimizada para ajuste fino eficiente de modelos abertos, incluindo suporte a quantização e integração com PEFT/LoRA (WOLF et al., 2020; Hugging Face, 2024).

O modelo base foi carregado por meio de `FastLanguageModel.from_pretrained`, com `load_in_4bit=True`, de modo a reduzir o consumo de memória durante o treinamento, seguindo o princípio do QLoRA, que combina o congelamento do modelo base quantizado em 4 bits com o treinamento de adaptadores LoRA, permitindo ajuste fino eficiente mesmo em GPUs com memória limitada e com baixo custo computacional (DETTMERS et al., 2023). Além disso, o comprimento máximo de sequência foi definido como `1024`, equilibrando contexto disponível e consumo de memória.

Em seguida, foi aplicada a técnica LoRA (Low-Rank Adaptation) por meio de `get_peft_model`, com os hiperparâmetros `r=16`, `lora_alpha=32`, `lora_dropout=0.05` e `bias="none"`. A LoRA reduz o número de parâmetros treináveis ao inserir matrizes de baixa dimensão nas camadas do Transformer, preservando os pesos originais do modelo

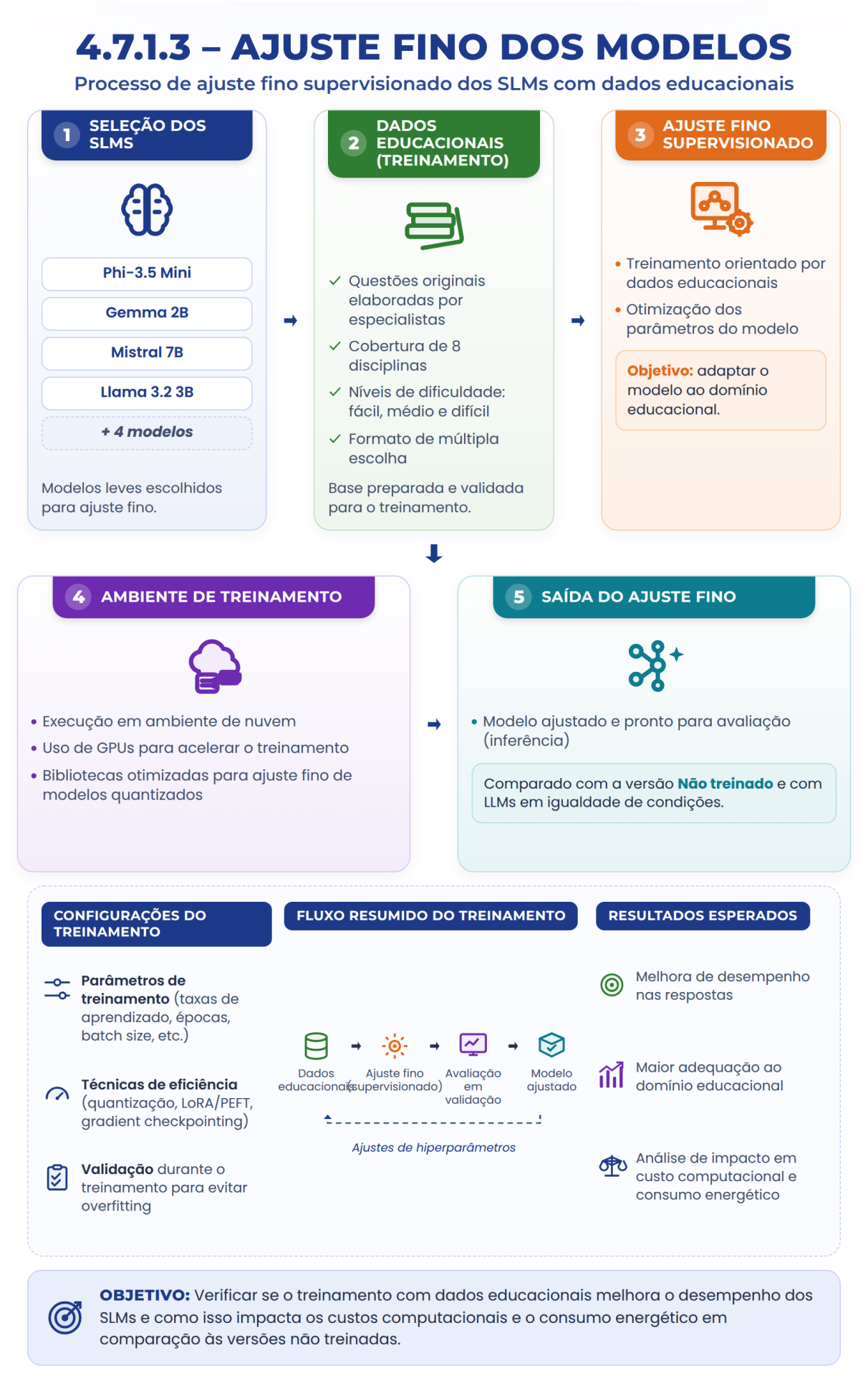


Figura 4 – Visão geral do processo de ajuste fino supervisionado aplicado aos SLMs com dados educacionais.

base e diminuindo o custo de ajuste fino (HU et al., 2022). Para este experimento, foram selecionados os módulos `q_proj`, `k_proj`, `v_proj` e `o_proj` por corresponderem às projeções de atenção, que têm papel central na adaptação do comportamento do modelo. Também foi habilitado `gradient_checkpointing`, técnica que reduz o uso de memória ao recomputar ativações durante a retropropagação.

O conjunto de treinamento foi armazenado em formato JSONL e carregado através do `load_dataset`. Cada amostra foi convertida para o formato instrucional contendo três campos: *Instruction*, *Input* e *Response*. Essa formatação é apropriada para treinamento supervisionado, pois apresenta ao modelo uma instrução clara, um contexto adicional quando necessário e a resposta esperada como saída alvo (LHOEST et al., 2021).

O conjunto de dados utilizado no treinamento foi obtido através da junção de 3 conjuntos de dados públicos que possuíam um conjunto de dados de referência para respostas de perguntas cruzadas e multilíngues para exames do ensino médio. Entre eles estão: **ScienceQA**, conjunto de dados coletado a partir dos currículos de ciências do ensino fundamental e médio, e contém 21.208 conteúdos multimodais de múltipla escolha envolvendo áreas de Biologia, Física, Geografia, História e derivados (LU et al., 2022); **RACE**, um conjunto de dados coletado a partir dos exames de inglês para alunos chineses do ensino fundamental e médio na faixa etária entre 12 e 18 anos, com perguntas geradas por especialistas humanos e aborda uma variedade de tópicos cuidadosamente elaborados para avaliar a capacidade dos alunos de entender e raciocinar (LAI et al., 2017); e **GSM8K**, conjunto de dados focado em problemas de matemática de alta qualidade para o ensino fundamental, que demandam de 2 e 8 etapas para serem resolvidos e envolvem principalmente a realização de uma sequência de cálculos utilizando operações aritméticas básicas (COBBE et al., 2021). Das 3 bases de dados, foram selecionadas aleatoriamente 42 mil perguntas para compor a base de dados final utilizada no ajuste fino dos modelos.

O treinamento foi então conduzido com o `SFTTrainer`, que encapsula a lógica de ajuste fino supervisionado e utiliza, internamente, os argumentos de treinamento definidos em `TrainingArguments` (Hugging Face, 2024).

A configuração de treinamento foi definida com `per_device_train_batch_size=1` e `gradient_accumulation_steps=16`, o que resulta em um lote efetivo maior sem exigir aumento proporcional da memória da GPU. Foram utilizados ainda `learning_rate=2e-4`, `num_train_epochs=1`, `fp16=True`, `warmup_steps=100`, `lr_scheduler_type="cosine"`, `optim="adamw_torch"` e `max_grad_norm=0.3`.

Durante a execução, foram configurados `logging_steps=50`, `save_steps=50` e `save_total_limit=3`, de modo a registrar o progresso do treinamento e manter apenas os checkpoints mais recentes. Antes do início de cada execução, o script verifica a existência do último checkpoint por meio de `get_last_checkpoint`; caso ele seja encontrado, o treinamento é retomado automaticamente a partir desse ponto, o que evita a perda de

progresso em caso de interrupção da sessão no Colab. Ao final do processo, o modelo ajustado foi salvo como adaptador LoRA, preservando a separação entre o modelo base e os pesos adicionados durante o ajuste fino.

Após o treinamento, o adaptador LoRA foi incorporado ao modelo base por meio do processo de *merge*, gerando uma versão consolidada para inferência. Essa etapa transforma o modelo ajustado em uma unidade única, sem dependência externa do adaptador. Em seguida, o modelo foi convertido para o formato GGUF, formato amplamente utilizado para distribuição e execução local de modelos de linguagem, facilitando carregamento rápido e execução do modelo fora do ambiente de treinamento, através da conversão utilizando o *llama.cpp*, projetado para armazenamento e inferência eficientes em modelos quantizados (GERGANOV, 2023).

4.7.1.4 Engenharia de prompts e inferência

Nesta etapa, implementou-se a engenharia de *prompts* para otimizar o desempenho das SLMs e garantir a configuração fixa dos parâmetros de geração. Este processo teve como finalidade principal a garantia de padronização de inferência, com os parâmetros fixados para permitir a comparabilidade na função de decisão e reduzir os vieses de geração.

Cada prompt incorporou quatro componentes principais: instrução, contexto, dados de entrada e indicador de saída, garantindo que o modelo compreendesse a solicitação e fornecesse respostas consistentes (WHITE et al., 2023).

Cada requisição enviada ao modelo incluiu:

- Instruções iniciais ao assistente, definindo a tarefa e a forma de resposta;
- Pergunta a ser respondida e opções de resposta.

Além disso, foram configurados parâmetros de amostragem dos modelos, incluindo Temperatura, *Top-k*, *Top-p* e *Min-p*, de modo a padronizar a geração de respostas entre todos os modelos avaliados.

Por meio de seu estudo, Sachete et al. (2025) constatou em uma análise de vários modelos em um contexto educacional que: temperaturas mais baixas favorecem respostas diretas, com alta fidelidade lexical e menor custo computacional; valores intermediários como 0,5 se destacam por equilibrar diversidade e coerência, sendo particularmente eficazes em tarefas que exigem inferência e interpretação; em temperaturas mais elevadas, modelos mais robustos demonstraram maior inventividade sem comprometer a coerência, enquanto modelos menores revelaram instabilidade semântica.

Através desses dados, foi decidido que a temperatura a ser utilizada nos modelos como padrão irá ser de 0,3, mantendo as respostas mais determinísticas, reduzindo criatividade excessiva e risco de invenções, mantendo também alto desempenho.

De acordo com Hewitt, Manning e Liang (2022), os valores de *top-p* entre 0,7 e 0,95 são os mais eficazes em tarefas educacionais, garantindo que o modelo mantenha coerência sem restringir excessivamente o espaço de geração. Neste contexto, o *top-p* elevado demais pode levar à inserção de termos imprecisos ou irrelevantes, enquanto valores muito baixos podem limitar a expressividade da resposta.

Diante disso, definiu-se o valor de *top-p* = 0,8 como padrão, permitindo um bom balanço entre clareza e riqueza de explicação, reduzindo a chance de alucinações e com flexibilidade para frases naturais, mantendo a coerência.

Estudos como o de Trinh e Dai (2022) mostram que *top-k* sozinho pode ser menos eficaz que *top-p*, já que impõe um limite fixo mesmo quando a distribuição é muito desigual. Contudo, a combinação de *top-k* moderado com *top-p* melhora a robustez contra tokens improváveis.

Para este trabalho, optou-se por um valor de *top-k* = 40, garantindo que o modelo mantenha naturalidade no discurso, mas sem explorar tokens com baixíssima probabilidade.

Conforme demonstrado em Nguyen et al. (2025), esse método é especialmente útil para reduzir incoerências sem sacrificar variedade, funcionando como complemento a *top-p*. Em tarefas educacionais, mostrou-se eficaz para manter precisão sem comprometer a clareza explicativa.

Neste trabalho, foi adotado *min-p* = 0,1, garantindo respostas de alta fidelidade lexical, reduzindo significativamente a inclusão de tokens sem relevância para o contexto.

A fim de evitar erro de looping na execução dos modelos, foi definido o tamanho máximo de tokens de saída por resposta como 1000.

4.7.2 Fase de Intervenção

Após a definição de todos os componentes necessários para a condução do experimento, iniciou-se a etapa de execução padronizada prática dos modelos para a coleta de métricas para alimentar a decisão.

Para o carregamento e execução de cada modelo, foi utilizado o software *LM Studio*, uma ferramenta que permite a execução local de modelos de linguagem, disponibilizando-os por meio de uma API compatível com o padrão da OpenAI. Essa abordagem proporciona maior controle sobre a configuração e a inferência dos modelos, além de eliminar custos associados ao uso de APIs comerciais, consolidando-se como uma solução estratégica para prototipagem e realização de experimentos intensivos (JUNIOR et al., 2025).

Para a medição do consumo energético, utilizou-se o software *HWiNFO*, uma ferramenta abrangente de monitoramento e diagnóstico de hardware, capaz de fornecer informações em tempo real sobre diversos parâmetros do sistema, como consumo de energia, potência, utilização da CPU e GPU (AHMED; IQBAL; JAMIL, 2024). Adicionalmente, a ferramenta permite o registro contínuo desses dados em intervalos de 0,5 segundos, possibilitando a coleta detalhada de métricas essenciais para os critérios do *framework*.

Dessa forma, ao iniciar o script pelo terminal de comando, era desencadeado todo o fluxo de execução. Esse processo incluía a inicialização do modelo no LM Studio, a recuperação de todas as perguntas — originais e não originais — no banco de dados, sua execução individual no modelo e o registro de todas as informações de processamento no banco de dados. Ao término de cada ciclo, o script era executado novamente para o próximo modelo.

Com isso, a cada interação são gerados e armazenados os dados para a utilização futura da função de decisão.

O hardware utilizado para este experimento foi um computador desktop com as seguintes especificações:

- Processador: AMD Ryzen 9 5900X, 12 núcleos, 24 threads, clock de 3.7 GHz
- Placa de Vídeo: AMD Radeon RX 6700 XT, 12 GB GDDR6
- Memória RAM: 32GB (4x8) 3600MHz DDR4
- Placa Mãe: Asus TUF Gaming B550M-Plus
- Sistema Operacional: Windows 11 Pro, Versão 25H2, SO 26200.8037

O mesmo hardware foi utilizado durante todo o processo de obtenção dos dados, com isso essa padronização do hardware garante a comparabilidade. Cabe ressaltar que a capacidade de 12 GB de VRAM da GPU utilizada foi suficiente para executar integralmente os modelos de até aproximadamente 12 bilhões de parâmetros na quantização Q8. Modelos maiores (GPT OSS 20B, Gemma 3 27B, Qwen 3 30B e LLaMA 3.1 70B) excederam esse limite e recorreram ao mecanismo de *offloading* para CPU/RAM oferecido pelo *llama.cpp*, no qual parte das camadas do modelo é processada pela CPU utilizando memória principal. Esse comportamento explica a inversão observada na distribuição de carga energética entre CPU e GPU para os modelos de grande porte, discutida na análise descritiva. Trata-se de uma condição realista para os cenários educacionais de infraestrutura restrita explorados pelo *framework*, mas que impõe cautela na comparação absoluta de latência e consumo entre modelos que couberam integralmente na VRAM e modelos que sofreram *offloading*.

Após a geração das saídas por cada modelo, realizou-se uma validação minuciosa dos resultados. Durante a execução do script, as respostas que apresentavam correspondência

exata com o gabarito foram computadas automaticamente como acertos. Contudo, divergências textuais mínimas ou informações adicionais impediam a classificação automatizada. Por esse motivo, os autores conduziram uma revisão manual posterior de todas as respostas para avaliar rigorosamente a sua correção.

A repetibilidade das medições foi endereçada por dois mecanismos complementares. Primeiro, a padronização integral do ambiente de execução: o mesmo hardware, o mesmo software de inferência, a mesma configuração térmica e a ausência de processos concorrentes de usuário durante as execuções garantem que medições sucessivas da mesma pergunta com o mesmo modelo produziriam valores dentro de faixas de variação próximas. Segundo, a amostragem contínua do HWiNFO a intervalos de 0,5 segundos ao longo de cada requisição fornece uma medida integrada da energia consumida, o que reduz sensibilidade a flutuações instantâneas. Não obstante, dada a natureza estocástica da inferência em modelos de linguagem e as variações naturais de potência elétrica ao nível do sistema operacional, os valores absolutos de energia e latência devem ser interpretados como estimativas com margem de variação inerente.

O consumo energético observado corresponde à leitura direta dos sensores do sistema durante a execução da inferência, o que inclui, além da carga do modelo em si, o consumo basal do sistema operacional e de processos mínimos de fundo inerentes ao ambiente Windows. Para mitigar a influência desses fatores, todas as execuções foram realizadas sob condições padronizadas: ausência de aplicações abertas além do LM Studio e do HWiNFO, sem execução simultânea de outros processos intensivos e sem estresse térmico prévio da GPU. Ainda assim, reconhece-se que o consumo absoluto observado incorpora componentes não atribuíveis exclusivamente ao modelo avaliado. A comparação entre modelos permanece válida por essa componente basal ser aproximadamente constante entre execuções realizadas em condições equivalentes.

4.7.3 Fase de Análise

A fase de análise foi planejada com um duplo propósito: responder às hipóteses formuladas ao longo da pesquisa sobre o comportamento dos modelos e validar empiricamente o *framework* proposto para seleção multicritério de modelos de linguagem em contextos educacionais. A análise concentrou-se na investigação do comportamento dos modelos frente às diferentes variáveis experimentais, considerando simultaneamente aspectos relacionados à acurácia, latência, custo computacional e consumo energético, sendo representada em todas as suas etapas pela Figura 5.

Ao invés de adotar uma leitura baseada exclusivamente em métricas isoladas, esta pesquisa assume uma perspectiva multidimensional, segundo a qual o desempenho de modelos de linguagem depende do equilíbrio entre critérios operacionais, computacionais e pedagógicos. Dessa forma, a etapa analítica foi estruturada para identificar não apenas



Figura 5 – Visão geral da fase de análise.

diferenças globais entre modelos, mas também interações entre características do contexto experimental, como disciplina, nível de dificuldade e originalidade das questões.

Considerando que cada modelo foi submetido ao mesmo conjunto de questões distintas, com observações associadas ao mesmo modelo, à mesma disciplina, ao mesmo nível de dificuldade e ao mesmo tipo de questão, adotaram-se procedimentos estatísticos capazes de lidar com dependência entre observações e com diferentes naturezas de variáveis resposta. Com isso, a etapa analítica cumpre simultaneamente uma função avaliativa e uma função de refinamento do artefato.

4.7.3.1 Estrutura dos dados experimentais

Os dados experimentais foram obtidos a partir das execuções dos modelos de linguagem sobre conjuntos de questões educacionais organizadas por disciplina, nível de dificuldade e originalidade. Cada interação realizada pelos modelos gerou um conjunto de informações armazenadas no banco de dados da aplicação experimental, permitindo a construção de um conjunto estruturado de observações para análise estatística.

As informações registradas contemplaram:

- modelo utilizado;
- categoria do modelo (*SLM* ou *LLM*);
- condição do modelo (treinado ou não treinado);
- disciplina associada à questão;
- nível de dificuldade da questão;
- originalidade da questão;
- resposta produzida pelo modelo;
- acerto ou erro da resposta;
- quantidade de tokens utilizados;
- tempo de resposta (s);
- consumo energético (J) associado à execução.

4.7.3.2 Variáveis analisadas

A análise experimental foi conduzida considerando variáveis dependentes e independentes relacionadas ao comportamento dos modelos durante o processo de inferência.

As variáveis dependentes avaliadas foram:

- acurácia das respostas;
- tempo de resposta;
- quantidade de tokens utilizados;
- consumo energético.

A acurácia foi adotada como principal indicador de desempenho cognitivo dos modelos, por representar sua capacidade de responder corretamente às questões propostas. Já o tempo de resposta, a quantidade de tokens utilizados e o consumo energético foram empregados como indicadores de custo computacional e de eficiência operacional.

As variáveis independentes consideradas na análise foram:

- tipo de modelo (*SLM* ou *LLM*);
- modelo específico utilizado;
- condição de treinamento (treinado ou não treinado);
- disciplina;
- nível de dificuldade;
- originalidade da questão.

A definição dessas variáveis permitiu investigar não apenas o comportamento geral dos modelos, mas também a existência de interações entre fatores contextuais e métricas de desempenho.

4.7.3.3 Estratégia de análise estatística

A estratégia de análise estatística foi organizada em três níveis complementares:

1. análise descritiva das métricas coletadas;
2. análise inferencial das hipóteses propostas;
3. validação empírica do *framework* multicritério.

Inicialmente, foram realizadas análises descritivas das métricas obtidas durante os experimentos, incluindo medidas de tendência central, dispersão e distribuição dos dados. Em seguida, foram conduzidas análises inferenciais com o objetivo de investigar relações entre as variáveis experimentais e testar as hipóteses formuladas na pesquisa.

Dessa forma, foram utilizados modelos de Equações de Estimação Generalizadas (*GEE*) para variáveis binárias e medidas repetidas, modelos lineares de efeitos mistos para variáveis contínuas com estrutura agrupada, e o procedimento *TOST* (*Two One-Sided Tests*) para avaliação de equivalência prática entre grupos (LIANG; ZEGER, 1986; PINHEIRO; BATES, 2000; LAKENS, 2017). Variáveis contínuas assimétricas, como tempo de resposta, consumo de tokens e consumo energético, foram analisadas em escala logarítmica quando necessário, com o objetivo de reduzir a assimetria, estabilizar a variância e melhorar o ajuste dos modelos.

A variável acurácia é dicotômica, pois cada resposta pode ser classificada como correta ou incorreta. Por esse motivo, as análises envolvendo acurácia foram conduzidas por meio de *GEE* com distribuição binomial, técnica adequada para estimar efeitos populacionais em dados correlacionados, especialmente quando há medidas repetidas ou dependência entre observações (LIANG; ZEGER, 1986; HARRIS; YANG; HARDIN, 2012). As *GEE* permitem modelar respostas não independentes sem exigir a suposição de independência completa entre as observações, oferecendo estimativas robustas para efeitos médios populacionais em estudos com dados agrupados, repetidos ou correlacionados (LIANG; ZEGER, 1986; HARRIS; YANG; HARDIN, 2012).

Para as variáveis contínuas ou quantitativas associadas à eficiência computacional, como tempo de resposta, quantidade de tokens e consumo energético, foram utilizados modelos lineares de efeitos mistos ou *GEE*, conforme a natureza da variável e a estrutura de dependência observada nos dados (PINHEIRO; BATES, 2000). Os modelos de efeitos mistos são apropriados para dados agrupados ou longitudinais, pois permitem considerar fontes de variação associadas aos agrupamentos experimentais, como questões, execuções ou modelos, além dos efeitos fixos de interesse.

Quando necessário, variáveis com distribuição assimétrica, cauda longa ou heterocedasticidade foram transformadas em escala logarítmica. Essa estratégia contribui para estabilizar a variância, reduzir assimetrias e tornar as comparações entre grupos mais robustas (MANNING, 1998; CHANGYONG et al., 2014).

As análises estatísticas não foram conduzidas apenas para comparar modelos de forma isolada, mas para verificar se o *framework* é capaz de identificar, com base nas evidências observadas, o modelo mais adequado a cada cenário educacional. Nesse sentido, a análise das hipóteses foi organizada em quatro dimensões interpretativas: (i) desempenho por complexidade da tarefa; (ii) eficiência computacional; (iii) *trade-offs* entre desempenho e custo; e (iv) adequação contextual e validação da decisão produzida pelo *framework*. Essa organização permitiu interpretar os resultados de forma progressiva, da comparação direta entre modelos até a verificação de que a proposta multicritério é mais adequada do que escolhas baseadas em uma única métrica.

4.7.3.4 Análise das hipóteses

As hipóteses da pesquisa foram avaliadas por meio da análise conjunta das métricas coletadas durante os experimentos. A Tabela 10 apresenta a relação entre as hipóteses investigadas, as variáveis analisadas e os métodos estatísticos utilizados.

Tabela 10 – Relação entre hipóteses, variáveis e métodos estatísticos.

Hipótese	**Variável resposta**	**Variáveis explicativas**	**Método estatístico**
H1	Acurácia	Tipo de modelo × dificuldade	*TOST* (delta 0,05) + *GEE* binomial
H2	Acurácia	Tipo de modelo em questões difíceis	*GEE* binomial com contraste direcional
H3	Tempo de resposta	Tipo de modelo	Modelos de efeitos mistos ou *GEE*
H4	Consumo de tokens	Tipo de modelo	Modelos de efeitos mistos ou *GEE*
H5	Consumo energético	Tipo de modelo	Regressão OLS e *Mann-Whitney*
H6	Custo computacional	Porte do modelo + termo quadrático	Modelo com termo linear e quadrático
H7	Custo energético	Porte do modelo + termo quadrático	Modelo com termo linear e quadrático
H8	Acurácia	Tokens × disciplina × dificuldade × originalidade	*GEE* binomial com interações
H9	Acurácia, tempo, tokens e energia	Condição de treinamento	*GEE* binomial + modelos de efeitos mistos
H10	Desempenho contextual	Modelo × disciplina × dificuldade × originalidade	Análise de interações e perfis de desempenho
H11	Adequação ao cenário	Critérios multicritério	*Profile Matching* + comparação com decisão por métrica isolada

4.7.3.5 Dimensão 1: Acurácia por Complexidade

A primeira dimensão analítica concentrou-se na avaliação se os modelos mantêm acurácia competitiva à medida que a complexidade das tarefas aumenta. Essa dimensão está diretamente relacionada às hipóteses **H1** e **H2**.

Para a hipótese **H1**, que afirma que *SLMs* de porte médio apresentam desempenho equivalente a *LLMs* em tarefas de nível Fácil e Médio, a análise foi conduzida por meio do procedimento *TOST* para avaliação de equivalência prática entre os grupos. Diferentemente dos testes tradicionais de diferença estatística, cujo objetivo é identificar se existe diferença

significativa entre médias, o *TOST* com delta de 0,05 em pontos percentuais permite verificar se a diferença observada permanece dentro de uma margem previamente definida como aceitável do ponto de vista aplicado (LAKENS, 2017). Nesse contexto, a preocupação central não é apenas demonstrar ausência de diferença estatística, mas avaliar se eventuais discrepâncias entre *SLMs* e *LLMs* são suficientemente pequenas para que ambos possam ser considerados equivalentes em cenários educacionais de menor complexidade.

Complementarmente, a acurácia foi modelada por meio de *GEE* com distribuição binomial, considerando que a variável resposta é dicotômica (acerto ou erro) e que as observações apresentam dependência decorrente de execuções repetidas dos mesmos modelos em diferentes conjuntos de questões (LIANG; ZEGER, 1986). O uso de *GEE* é apropriado em contextos com medidas correlacionadas, pois permite estimar efeitos populacionais de forma robusta sem assumir independência completa entre as observações (HARRIS; YANG; HARDIN, 2012). Dessa forma, a análise permite verificar simultaneamente a equivalência prática entre os grupos e o comportamento dos modelos em cenários de menor exigência cognitiva.

Para a hipótese **H2**, segundo a qual *LLMs* apresentam desempenho superior em tarefas de nível Difícil, foi empregada uma modelagem por meio de *GEE* binomial com contraste direcional. Essa escolha se justifica porque a hipótese possui sentido explícito de superioridade, isto é, espera-se que os *LLMs* apresentem maior probabilidade de acerto em questões que exigem maior capacidade de raciocínio, generalização e integração de informações. Assim, a análise não se limita à verificação de diferenças médias, mas investiga se a superioridade dos *LLMs* emerge de forma consistente nos cenários em que a complexidade é mais elevada (LIANG; ZEGER, 1986; PINHEIRO; BATES, 2000).

4.7.3.6 Dimensão 2: Eficiência Computacional

A segunda dimensão avaliou o comportamento dos modelos em termos de eficiência computacional, considerando três aspectos: tempo de resposta, consumo de tokens e consumo energético. Essa dimensão está relacionada às hipóteses **H3**, **H4**, **H5**, **H6** e **H7**.

As hipóteses **H3**, **H4** e **H5**, que comparam tempo de resposta, consumo de tokens e consumo energético entre *SLMs* e *LLMs*, foram testadas por meio de modelos de efeitos mistos ou *GEE*, conforme a natureza de cada variável e a dependência entre as observações. O uso de *GEE* se justifica quando há medidas repetidas ou correlacionadas, enquanto os modelos de efeitos mistos são apropriados para dados agrupados ou longitudinais (LIANG; ZEGER, 1986; PINHEIRO; BATES, 2000). O tempo de resposta foi analisado em escala logarítmica, e as variáveis de tokens e energia também foram transformadas quando necessário, especialmente em casos de cauda longa ou heterocedasticidade, o que contribui para estabilizar a variância e tornar a comparação entre os grupos mais robusta (MANNING, 1998; CHANGYONG et al., 2014). Dessa forma, foi possível avaliar, com base

nos dados do experimento, se os modelos menores apresentam de fato melhor eficiência operacional.

As hipóteses **H6** e **H7** investigaram se o custo computacional e o custo energético aumentam de forma não linear com o porte do modelo. Para essa avaliação, foram comparadas especificações estatísticas com efeito linear do tamanho do modelo e especificações que incluíam, adicionalmente, um termo quadrático. Essa estratégia permitiu verificar se o aumento de parâmetros leva a um crescimento proporcional ou acelerado do custo, oferecendo uma leitura mais realista da relação entre escala do modelo e consumo de recursos (PINHEIRO; BATES, 2000). Essa etapa é central para a interpretação do *framework*, pois evidencia se ganhos marginais de desempenho passam a ser acompanhados por custos crescentes de forma desproporcional.

4.7.3.7 Dimensão 3: *Trade-offs* entre Desempenho e Custo

A terceira dimensão examinou a relação entre desempenho e custo, buscando identificar em que condições o aumento de acurácia ocorre junto com maior custo computacional ou energético. Essa dimensão está associada às hipóteses **H8** e **H9**.

A hipótese **H8** propõe que existem *trade-offs* significativos entre acurácia e custo computacional, sendo considerados modelos com termos de interação entre acurácia (A), tempo (L), custo computacional (C) e consumo energético (E), incorporando ainda disciplina, nível de dificuldade e originalidade da questão como variáveis moderadoras. A inclusão dessas interações é essencial para verificar se a relação entre desempenho e custo permanece estável em diferentes cenários ou se varia conforme o domínio avaliado, a complexidade da tarefa e a natureza da questão. Assim, a análise permite identificar situações em que maior custo se traduz em ganhos reais de desempenho, bem como cenários em que esse custo adicional não produz melhora relevante.

A hipótese **H9**, segundo a qual modelos treinados apresentam ganho de desempenho em relação aos não treinados, embora esse ganho possa implicar variações no custo computacional e energético, foi avaliada por meio da comparação entre versões treinadas e não treinadas dos mesmos modelos, utilizando *GEE* binomial para acurácia e modelos lineares de efeitos mistos para tempo, tokens e energia. Essa comparação permite isolar o efeito do treinamento e verificar se o ganho obtido compensa o custo adicional associado ao processo de adaptação (LIANG; ZEGER, 1986; PINHEIRO; BATES, 2000). Quando apropriado, foram realizados contrastes entre variantes base e treinadas com escala semelhante de parâmetros, de modo a reduzir a influência de diferenças arquiteturais, tamanho do modelo e estratégias de tokenização. Esse procedimento permite isolar com maior precisão o efeito do treinamento sobre as métricas avaliadas.

4.7.3.8 Dimensão 4: Adequação Contextual e Decisão

A última dimensão atua como a validação final do próprio artefato construído, correspondendo às hipóteses **H10** e **H11**. Seu objetivo é demonstrar que a seleção de modelos baseada na aproximação de perfis (*Profile Matching*) é mais realista e útil do que olhar para as métricas isoladamente.

A hipótese **H10**, que estabelece que não existe um modelo universalmente superior em todos os cenários, foi testada por meio da análise das interações entre tipo de modelo, disciplina, nível de dificuldade e originalidade da questão. A presença de efeitos de interação significativos indica que o desempenho dos modelos não é homogêneo entre contextos, reforçando a necessidade de uma decisão contextualizada. Além disso, foram observados os perfis de desempenho por cenário, identificando eventuais reversões de *ranking* entre modelos, o que constitui evidência adicional de que a escolha do modelo depende da natureza da tarefa e não apenas de sua escala.

Por fim, a validação empírica do *framework* (**H11**) ocorreu pela simulação do seu mecanismo matemático. Diferentes perfis de uso (cenários alvo) foram criados estipulando faixas desejadas (Baixa, Média, Alta) para cada critério, considerando combinações distintas de exigências relacionadas à acurácia, tempo de resposta, custo computacional e eficiência energética. Os resultados empíricos obtidos durante os experimentos foram normalizados e utilizados como entrada da função de utilidade do *framework*. A partir disso, o algoritmo calculou a distância matemática entre as exigências do cenário e o desempenho real de todos os modelos avaliados, permitindo gerar recomendações contextualizadas de seleção.

A validação empírica do *framework* consistiu em comparar as recomendações produzidas pela função multicritério com estratégias de decisão baseadas em métricas isoladas. Essa comparação buscou verificar se a abordagem baseada em aproximação de perfis é capaz de produzir decisões mais adequadas às restrições reais dos contextos educacionais analisados.

Nesse sentido, a análise não se restringiu à identificação do modelo com maior acurácia, mas considerou o equilíbrio entre desempenho, custo computacional, tempo de resposta e consumo energético. Com isso, a etapa analítica buscou demonstrar que a seleção contextualizada de modelos pode produzir resultados mais coerentes com as necessidades operacionais e pedagógicas dos ambientes educacionais.

As análises estatísticas foram conduzidas considerando nível de significância de 5% ($p < 0{,}05$). Adicionalmente, foram avaliados indicadores de ajuste e qualidade dos modelos estatísticos, buscando garantir consistência e adequação das inferências realizadas.

4.7.3.9 Síntese da etapa analítica

De modo geral, a fase de análise foi concebida para transformar os dados experimentais obtidos durante a execução dos modelos em evidências quantitativas capazes de sustentar a validação do *framework* proposto. A integração entre análise estatística, avaliação multicritério e aproximação de perfis permitiu investigar, de maneira simultânea, desempenho, custo computacional, eficiência energética e adequação contextual, verificando não apenas se determinados modelos apresentam melhor desempenho em métricas específicas, mas principalmente se o *framework* proposto é capaz de operacionalizar os *trade-offs* observados e transformar dados experimentais em suporte estruturado à tomada de decisão em ambientes educacionais.

Essa abordagem reforça o caráter aplicado da pesquisa e contribui para a construção de um mecanismo estruturado de apoio à decisão para seleção de modelos de linguagem em contextos educacionais. Além disso, a etapa analítica permitiu operacionalizar os *trade-offs* existentes entre desempenho cognitivo, eficiência computacional e sustentabilidade, alinhando-se aos princípios da *Green AI* e às demandas contemporâneas de uso responsável de Inteligência Artificial na educação.

5 Resultados e Discussão

Nesta seção, são apresentados os resultados referentes à avaliação experimental, à confirmação do *framework* proposto e aos testes de hipóteses realizados.

5.1 Análise Experimental dos Modelos

Nesta subseção, são apresentados e discutidos os resultados obtidos na avaliação experimental dos modelos. As análises foram organizadas de acordo com as métricas de Acurácia, Latência, Custo Computacional e Consumo Energético.

5.1.1 Análise Descritiva da Acurácia por Modelo

A Tabela 11 apresenta os resultados descritivos de acurácia obtidos pelos modelos avaliados.

De maneira geral, observou-se que a acurácia está fortemente relacionada com o tamanho dos modelos. Os maiores índices de acurácia foram obtidos pelos modelos de grande porte, especificamente em suas versões não treinadas, com grande destaque para *Qwen 3 30B*, *LLaMA 3.1 70B*, *GPT OSS 20B* e *Gemma 3 27B*. Entre os modelos médios, *Qwen 2.5 7B*, *Qwen 3 4B* e *Phi 3.5 mini 3.8B* obtiveram destaque, superando outros modelos de mesma categoria (como *LLaMA 3.1 8B* e *Mistral 7B*) que possuem uma maior quantidade de parâmetros. Em contrapartida, os modelos pequenos (*LLaMA 3.2 1B*, *StableLM 2 1.6B* e *Gemma 2 2B*) registraram as menores acurácias do estudo, não ultrapassando a faixa dos 30% de acurácia.

Ao analisar o impacto direto do treinamento, observou-se que o efeito não foi uniforme. Para modelos pequenos, como *LLaMA 3.2 1B*, *StableLM 2 1.6B* e *LLaMA 3.1 8B*, as versões treinadas apresentaram ganhos consistentes em ambas as bases de questões. No entanto, para os modelos intermediários que já possuíam alta performance (*Qwen 2.5 7B*, *Qwen 3 4B* e *Phi 3.5 mini 3.8B*), o treinamento tendeu a elevar levemente os resultados nas questões não originais, mas gerou quedas perceptíveis de acurácia nas questões originais. No modelo *Mistral 7B* houve uma maior regressão, em que a versão treinada sofreu grande perda de acurácia em relação à versão base. Esse comportamento sugere que, para arquiteturas já otimizadas, o ajuste fino aos dados de treino pode ter gerado um sobreajuste (*overfitting*), prejudicando a capacidade dos modelos de responder às questões.

A natureza das questões também demonstrou ser um fator determinante na porcentagem de acerto, em que a ampla maioria dos modelos apresentou um desempenho

superior nas questões não originais. Em grandes modelos, a queda de desempenho nas questões originais foi expressiva, como no modelo *LLaMA 3.1 70B*, que caiu de 80,83% para 66,67%. Esse padrão observado entre os modelos sugere que as questões originais impõem um desafio maior de raciocínio ou que as questões não originais possam ter sido utilizadas no pré-treinamento desses modelos.

Tabela 11 – Acurácia por modelo em questões

Modelo	**Acurácia QNO**	**Acurácia QO**
LLaMA 3.2 1B	6,67%	17,50%
LLaMA 3.2 1B T	20,00%	20,00%
StableLM 2 1.6B	19,17%	20,83%
StableLM 2 1.6B T	30,83%	26,67%
Gemma 2 2B	20,83%	24,17%
Gemma 2 2B T	20,83%	20,83%
Phi 3.5 mini 3.8B	63,33%	56,67%
Phi 3.5 mini 3.8B T	64,17%	54,17%
Qwen 3 4B	65,83%	69,17%
Qwen 3 4B T	69,17%	65,00%
Mistral 7B	57,50%	47,50%
Mistral 7B T	44,17%	38,33%
Qwen 2.5 7B	75,83%	67,50%
Qwen 2.5 7B T	76,67%	55,83%
LLaMA 3.1 8B	42,50%	42,50%
LLaMA 3.1 8B T	55,00%	50,00%
Gemma 3 12B	76,67%	63,33%
GPT OSS 20B	80,00%	66,67%
Gemma 3 27B	78,33%	67,50%
Qwen 3 30B	80,83%	70,83%
LLaMA 3.1 70B	80,83%	66,67%

Nota: QNO refere-se a Questão Não Original, enquanto QO refere-se a Questão Original.

A Figura 6 representa a acurácia por modelo em questões não originais e a Figura 7 em questões originais.

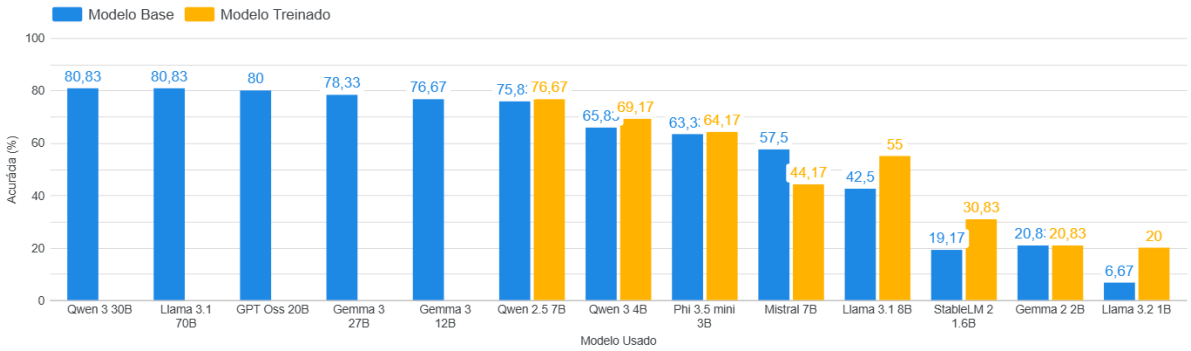


Figura 6 – Acurácia por Modelo em Questões Não Originais

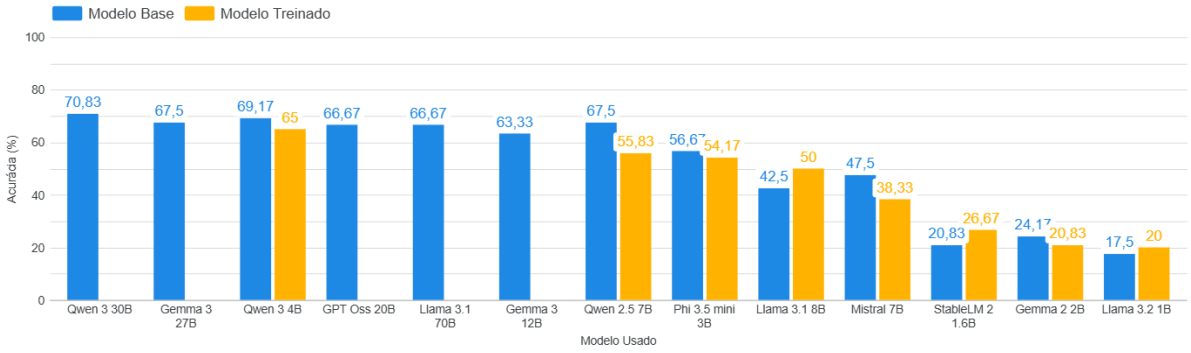


Figura 7 – Acurácia por Modelo em Questões Originais

5.1.2 Análise Descritiva da Latência por Modelo

A Tabela 12 apresenta os resultados descritivos do tempo médio de resposta (latência) obtidos pelos modelos avaliados.

De maneira geral, observou-se que a latência possui uma relação direta com o tamanho e a quantidade de parâmetros dos modelos, embora com exceções notáveis. Os maiores tempos de resposta foram registrados pelos grandes modelos, com destaque para o *LLaMA 3.1 70B*, que demandou até 25,42s, seguido por *Gemma 3 27B* e *GPT OSS 20B*, ambos superando a marca dos 10 segundos. No entanto, dentro dessa mesma categoria de grande porte, os modelos *Qwen 3 30B* e *Gemma 3 12B* demonstraram alta eficiência, mantendo médias abaixo dos 5 segundos. O destaque se concentra nos modelos de tamanho médio em suas versões não treinadas, que mesmo possuindo uma quantidade de parâmetros significativamente maior que os modelos pequenos apresentaram excelente eficiência temporal com tempos médios (questões não originais e originais) próximos ou abaixo de 1 segundo, como *Qwen 2.5 7B*, *Mistral 7B*, *Phi 3.5 mini 3.8B* e *LLaMA 3.1 8B*.

Ao analisar o impacto direto do treinamento, observou-se que o efeito causou um aumento expressivo no tempo de resposta na grande maioria dos modelos. Em modelos como *LLaMA 3.1 8B*, *Gemma 2 2B*, *Phi 3.5 mini 3.8B* e *Mistral 7B*, as versões treinadas sofreram um acréscimo drástico na latência, saltando, por exemplo, de 1,69s para 12,94s e 0,68s para 7,55s. Esse comportamento sugere que o ajuste fino pode ter induzido respostas mais longas, detalhadas ou estruturadas, exigindo maior tempo de geração. Contudo, a família Qwen apresentou um comportamento inverso e otimizado, reduzindo seus tempos de resposta após o treinamento nos seus dois modelos, evidenciando que o ajuste pode ter priorizado a concisão ou a eficiência computacional.

O tipo de questão demonstrou exercer uma influência pontual, mas relevante, na velocidade de processamento dos modelos. Para a maior parte das estruturas de pequeno e médio porte, a diferença de tempo entre questões originais e não originais foi sutil, havendo baixa diferença de valores, aumentando ou diminuindo. No entanto, nos modelos grandes,

a complexidade das questões originais resultou em um aumento perceptível na latência. O principal caso ocorreu no *LLaMA 3.1 70B*, cujo tempo de resposta saltou de 19,63s nas questões não originais para 25,42s nas originais. Esse padrão, quando correlacionado à queda de acurácia observada anteriormente, reforça a hipótese de que as questões originais exigem um esforço computacional e de raciocínio sensivelmente maior por parte das redes de grande porte, possivelmente resultando na geração de cadeias de pensamento mais densas antes da entrega da resposta final.

Tabela 12 – Tempo médio de resposta dos modelos em questões

Modelo	**Latência QNO**	**Latência QO**
LLaMA 3.2 1B	1,53s	0,90s
LLaMA 3.2 1B T	4,19s	2,79s
StableLM 2 1.6B	1,75s	1,57s
StableLM 2 1.6B T	4,32s	3,78s
Gemma 2 2B	0,68s	0,53s
Gemma 2 2B T	7,55s	7,68s
Phi 3.5 mini 3.8B	1,34s	1,39s
Phi 3.5 mini 3.8B T	4,84s	5,49s
Qwen 3 4B	7,07s	7,23s
Qwen 3 4B T	6,00s	6,36s
Mistral 7B	1,17s	1,13s
Mistral 7B T	5,11s	5,39s
Qwen 2.5 7B	0,92s	0,80s
Qwen 2.5 7B T	0,67s	0,63s
LLaMA 3.1 8B	1,69s	1,31s
LLaMA 3.1 8B T	12,94s	12,43s
Gemma 3 12B	2,98s	2,99s
GPT OSS 20B	10,64s	11,01s
Gemma 3 27B	11,49s	12,81s
Qwen 3 30B	4,11s	4,39s
LLaMA 3.1 70B	19,63s	25,42s

Nota: QNO refere-se a Questão Não Original, enquanto QO refere-se a Questão Original.

A Figura 8 representa a latência por modelo em questões não originais e a Figura 9 em questões originais.

5.1.3 Análise Descritiva do Custo Computacional por Modelo

A Tabela 13 apresenta os resultados descritivos do custo computacional obtidos pelos modelos avaliados.

De maneira geral, observou-se que o custo computacional não apresentou uma relação direta com o tamanho dos modelos, apresentando um cenário de alta eficiência mesmo em modelos maiores. Ao contrário da tendência observada na latência, os modelos grandes registraram consumos baixos. O grande destaque é a família de modelos *Gemma*,

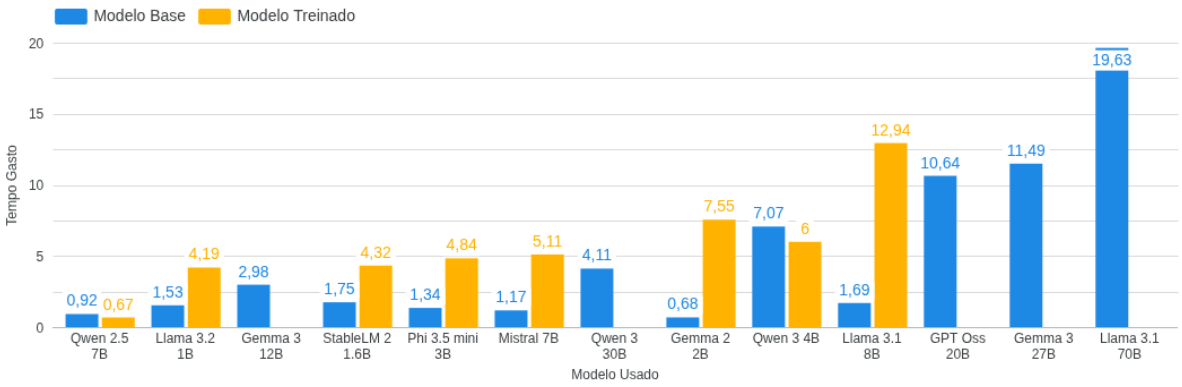


Figura 8 – Latência por Modelo em Questões Não Originais

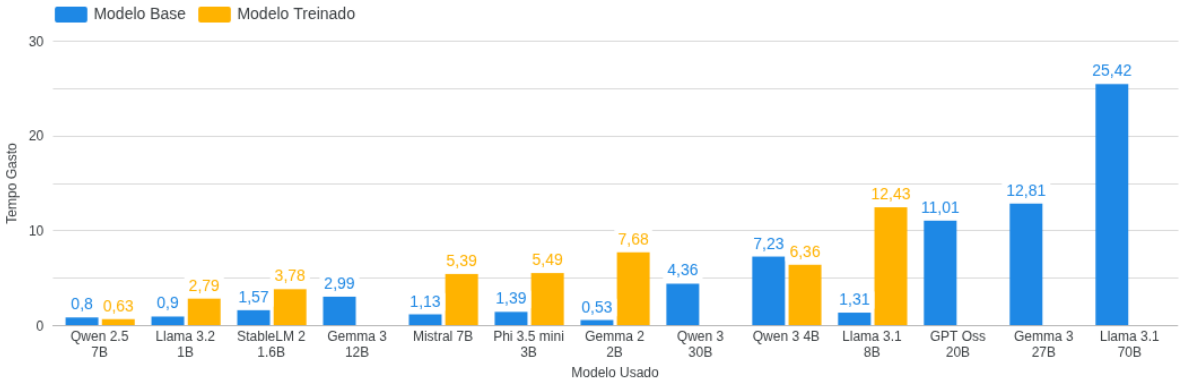


Figura 9 – Latência por Modelo em Questões Originais

na qual modelos grandes e pequenos apresentaram uma grande eficiência, com destaque para o *Gemma 3 12B* (média de 358,7), *Gemma 3 27B* (média de 359,38) e *Gemma 2 2B* (média de 369,55), que obtiveram os menores custos gerais do estudo em suas versões base.

Ao analisar o impacto direto do treinamento, observou-se que o efeito causou um aumento expressivo no custo computacional na grande maioria dos modelos. Em estruturas como *LLaMA 3.1 8B*, *Gemma 2 2B* e *StableLM 2 1.6B*, as versões treinadas sofreram um acréscimo drástico no consumo de tokens, que saltaram, por exemplo, de 455,49 para 888,98 e de 389,81 para 838,67. Esse comportamento valida perfeitamente os resultados de latência obtidos anteriormente, confirmando a hipótese de que o ajuste fino induziu essas redes a gerarem respostas muito mais longas, detalhadas ou estruturadas. Por outro lado, a família Qwen manteve seu comportamento inverso e otimizado; tanto o *Qwen 3 4B* quanto o *Qwen 2.5 7B* reduziram seus custos computacionais após o treinamento, ratificando que o ajuste nessas arquiteturas específicas priorizou a concisão e a economia de tokens.

O tipo de questão demonstrou exercer uma influência generalizada e altamente uniforme sobre o custo computacional de quase todas as estruturas avaliadas. Em todos os modelos, houve redução no consumo de tokens ao transicionar das questões não originais para as questões originais. Esse padrão de menor volume de tokens nas questões originais,

quando correlacionado à queda de acurácia e ao aumento de latência evidenciado nas seções anteriores, revela uma dinâmica analítica importante: diante do desafio e da complexidade superior impostos pelas questões originais, os modelos tendem a entregar respostas mais diretas, curtas ou compactas, despendendo um tempo de processamento significativamente maior por token gerado.

Tabela 13 – Custo Computacional por modelo em questões

Modelo	**Custo Computacional QNO**	**Custo Computacional QO**
LLaMA 3.2 1B	492,85	407,73
LLaMA 3.2 1B T	666,52	531,53
StableLM 2 1.6B	581,32	519,46
StableLM 2 1.6B T	858,78	772,11
Gemma 2 2B	389,81	349,29
Gemma 2 2B T	838,67	817,71
Phi 3.5 mini 3.8B	528,19	487,87
Phi 3.5 mini 3.8B T	733,50	731,91
Qwen 3 4B	826,00	797,04
Qwen 3 4B T	761,78	744,53
Mistral 7B	502,27	454,45
Mistral 7B T	626,46	590,62
Qwen 2.5 7B	427,78	382,96
Qwen 2.5 7B T	420,31	377,50
LLaMA 3.1 8B	455,49	399,53
LLaMA 3.1 8B T	888,98	829,33
Gemma 3 12B	375,31	342,09
GPT OSS 20B	699,88	679,34
Gemma 3 27B	375,20	343,57
Qwen 3 30B	428,66	388,21
LLaMA 3.1 70B	451,60	417,85

Nota: QNO refere-se a Questão Não Original, enquanto QO refere-se a Questão Original.

A Figura 10 representa o custo computacional por modelo em questões não originais e a Figura 11 em questões originais.

5.1.4 Análise Descritiva do Consumo Energético por Modelo

Esta seção apresenta os resultados referentes ao custo energético da inferência, detalhando o consumo total acumulado e as métricas de potência (média e máxima) demandadas pelos componentes de processamento (CPU e GPU).

5.1.4.1 Consumo energético

A Tabela 14 apresenta os resultados descritivos do consumo energético médio em joules obtidos pelos modelos avaliados.

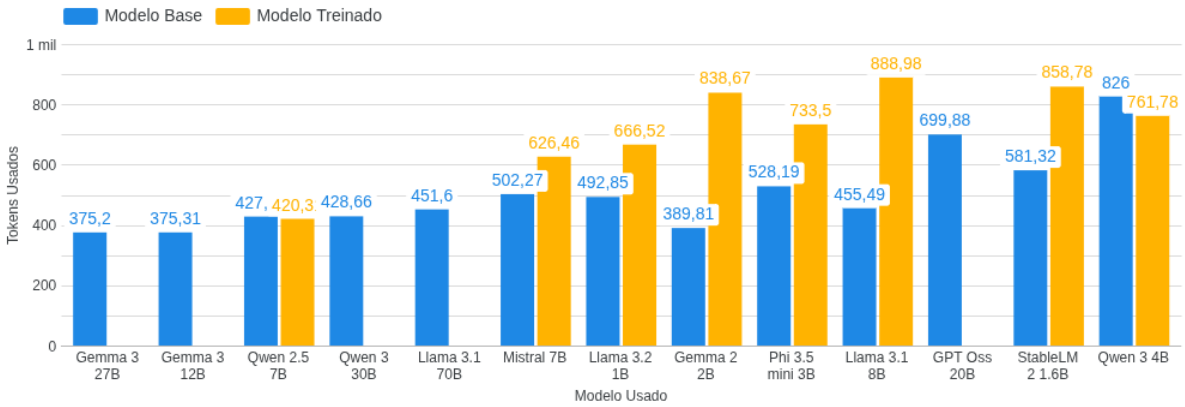


Figura 10 – Custo Computacional por Modelo em Questões Não Originais

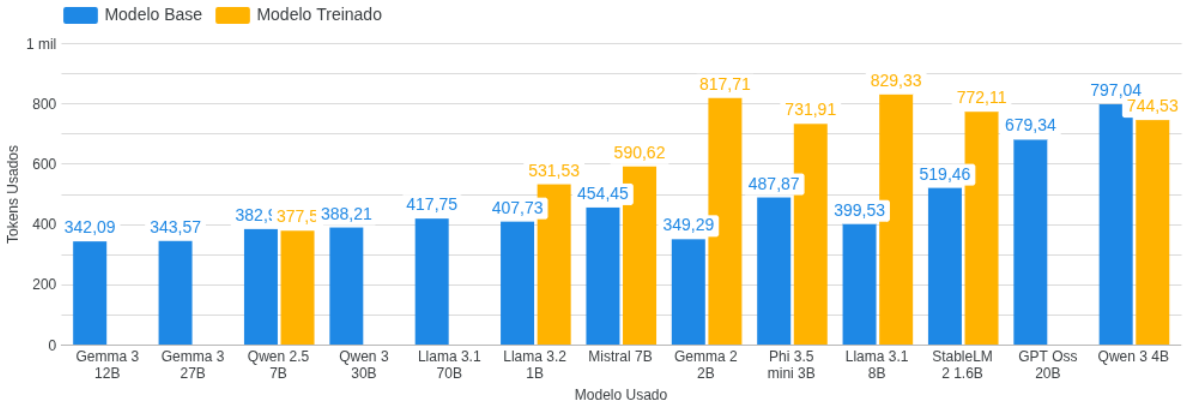


Figura 11 – Custo Computacional por Modelo em Questões Originais

De maneira geral, observou-se que o consumo energético possui uma relação direta com o porte e a quantidade de parâmetros dos modelos, embora com discrepâncias notáveis. Os maiores gastos energéticos globais foram registrados pelo maior modelo do experimento, o *LLaMA 3.1 70B*, que demandou até 5.091,76 J de energia total, seguido também de modelos grandes como *GPT OSS 20B* e *Gemma 3 27B*, ambos na faixa dos 2.000 J. Contudo, repetindo o padrão de alta eficiência observado nas seções anteriores, os modelos *Gemma 3 12B* e *Qwen 3 30B* conseguiram manter consumos surpreendentemente baixos para os seus tamanhos, não ultrapassando a marca dos 750 J. Entre os modelos médios, os destaques absolutos de economia energética se concentraram entre os modelos *Qwen 2.5 7B*, *Mistral 7B* e *Phi 3.5 mini 3.8B* em suas versões base, ocupando a faixa de 200J a 300J. Entre os modelos pequenos o grande destaque é o modelo *Gemma 2 2B* consumindo em média menos de 150J.

Ao analisar o impacto direto do treinamento, confirmou-se um aumento drástico e expressivo no consumo energético na grande maioria das estruturas. Modelos como *LLaMA 3.1 8B*, *Gemma 2 2B* e *Phi 3.5 mini 3.8B* viram suas demandas de joules multiplicarem severamente após o ajuste fino; o caso mais emblemático foi o do *LLaMA*

3.1 8B, que saltou de 404,10 J para 3.104,27 J, superando o gasto de modelos com maior quantidade de parâmetros. Esse comportamento alinha-se diretamente às evidências anteriores relacionadas à latência e custo computacional, validando a premissa de que o treinamento gerou respostas mais extensas e detalhadas, exigindo um tempo de execução prolongado que inflou o consumo total. Em contrapartida, a família Qwen consolidou sua assinatura de otimização inversa: tanto o *Qwen 3 4B* quanto o *Qwen 2.5 7B* reduziram seu consumo energético após passarem pelo treinamento, reforçando que o ajuste fino nessas arquiteturas priorizou respostas concisas e menor custo computacional.

O tipo de questão atuou de forma dicotômica entre as categorias de modelos, exercendo forte influência na distribuição de carga. Para os modelos grandes, a complexidade das questões originais resultou em uma elevação perceptível no consumo de joules, como visto em *LLaMA 3.1 70B*, no qual houve um acréscimo de 4.639,05 J para 5.091,76 J. Esse fenômeno demonstra que, embora essas redes tenham gerado menos tokens nas questões originais (conforme visto no custo computacional), o tempo de processamento estendido gerou um gasto de energia superior. Um detalhe crucial revelado pelos resultados consiste na divisão de consumo entre os componentes de hardware: enquanto nos modelos pequenos e médios a GPU dita quase a totalidade do consumo energético, em modelos grandes o consumo da CPU cresce exponencialmente, superando a GPU. Esse comportamento sugere que a movimentação de tensores e o gerenciamento de memória RAM para orquestrar bilhões de parâmetros impõem um esforço energético massivo ao processador principal.

Tabela 14 – Consumo energético médio em joules por modelo em questões

Modelo	Total QNO	CPU QNO	GPU QNO	Total QO	CPU QO	GPU QO
LLaMA 3.2 1B	278,62	111,65	166,97	159,5	64,5	95
LLaMA 3.2 1B T	749,61	301,41	448,2	513,5	215,59	297,91
StableLM 2 1.6B	383,73	124,41	259,32	342,08	112,46	229,62
StableLM 2 1.6B T	939,8	305,43	634,37	826,6	268,83	557,77
Gemma 2 2B	156,74	51,09	105,65	125,62	41,49	84,12
Gemma 2 2B T	1.651,15	554,49	1.096,66	1.692,88	578,17	1.114,7
Phi 3.5 mini 3.8B	306,54	71,75	234,8	325,72	82,54	243,19
Phi 3.5 mini 3.8B T	1.100,1	262,45	837,65	1.256,78	303,56	953,22
Qwen 3 4B	1.647,81	399,85	1.247,96	1.682,45	407,59	1.274,86
Qwen 3 4B T	1.403,33	344,47	1.058,88	1.489,28	364,52	1.124,76
Mistral 7B	279,52	72,8	206,72	270	72,65	197,35
Mistral 7B T	1.167,67	303,09	864,58	1.204,27	294,28	909,99
Qwen 2.5 7B	213,61	53,87	159,75	189,31	49,29	140,02
Qwen 2.5 7B T	170,72	48,34	122,38	149,01	38,34	110,67
LLaMA 3.1 8B	404,1	93,23	310,87	316,18	75,53	240,65
LLaMA 3.1 8B T	3.104,27	735,96	2.368,30	2.975,07	699,52	2.275,55
Gemma 3 12B	579,69	289,64	290,05	698,27	401,93	296,34
GPT OSS 20B	1.971,14	1.167	803,16	2.045,02	1.213,74	831,28
Gemma 3 27B	1.902,93	1.190,78	712,15	2.034,93	1.324,58	710,36
Qwen 3 30B	717,39	435,23	282,16	745,53	460,24	285,29
LLaMA 3.1 70B	4.639,05	2.574,54	2.064,51	5.091,76	2.958,62	2.133,14

Nota: QNO refere-se a Questão Não Original, enquanto QO refere-se a Questão Original.

A Figura 12 representa o consumo energético Médio por modelo em questões não

originais e a Figura 13 em questões originais.

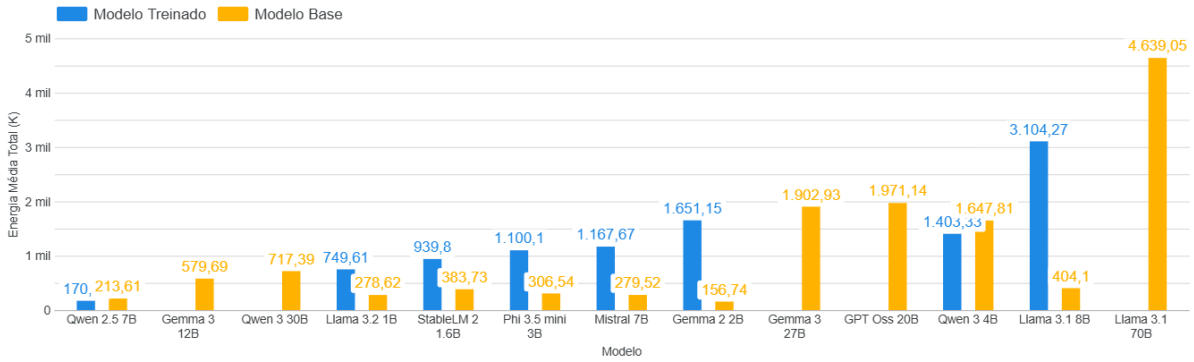


Figura 12 – Consumo Energético Médio em joules por Modelo em Questões Não Originais

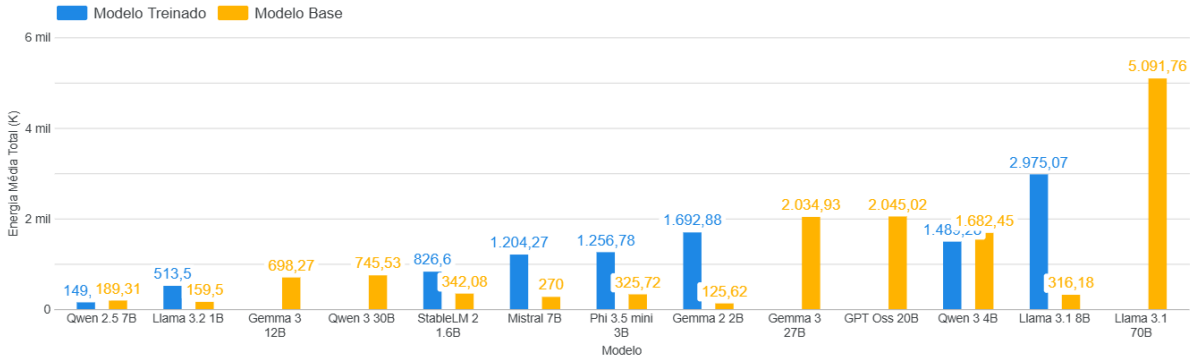


Figura 13 – Consumo Energético Médio em joules por Modelo em Questões Originais

5.1.4.2 Potência Média em watts

A Tabela 15 apresenta os resultados descritivos da potência média demandada, em watts, pelos modelos avaliados.

De maneira geral, observou-se que a potência média não apresentou uma correlação direta com o tamanho dos modelos, mantendo-se concentrada em uma faixa operacional bastante estável entre 140W e 242W para quase todas as arquiteturas. Diferente do comportamento visto no consumo energético total, os modelos de grande porte registraram demandas de potência contidas, com destaque para o *Gemma 3 27B* (165,35W) e o *Qwen 3 30B* (174,97W), que operaram com taxas inferiores às de modelos com menor quantidade de parâmetros. O pico de exigência de potência concentrou-se na categoria de tamanho médio, liderada pelo *LLaMA 3.1 8B* e pelo *Qwen 3 4B*, ambos operando de forma contínua acima de 235W.

Ao analisar o impacto direto do treinamento, constatou-se que o efeito foi praticamente nulo na alteração da potência consumida pela grande maioria das estruturas. Ao contrário das variações drásticas observadas na latência e na energia total, os modelos mantiveram

patamares de watts quase idênticos entre suas versões de base e treinadas. Esse comportamento evidencia que o processo de ajuste fino não altera a intensidade do esforço instantâneo do hardware, confirmando que o consumo energético observado anteriormente decorre exclusivamente do tempo estendido de execução (latência) e não de uma sobrecarga na taxa de potência operacional.

O tipo de questão também demonstrou exercer uma influência sutil, afetando de forma mais perceptível os modelos grandes. Enquanto os modelos pequenos e médios mantiveram variações irrelevantes de watts entre as bases, os modelos grandes registraram uma redução na potência média ao lidarem com as questões originais. Essa queda, combinada com a clara inversão na distribuição de carga na qual as CPUs dos grandes modelos passam a exigir mais watts do que as suas próprias GPUs, indica que o processamento de questões inéditas e complexas tende a desacelerar o ritmo de fluxo na GPU, concentrando o esforço de potência no gerenciamento lógico e de memória por parte da CPU.

Tabela 15 – Potência média demandada em watts por modelo em questões

Modelo	Total QNO	CPU QNO	GPU QNO	Total QO	CPU QO	GPU QO
LLaMA 3.2 1B	182,2	73,01	109,19	180,57	73,02	107,55
LLaMA 3.2 1B T	181,54	73	108,54	184,49	77,46	107,03
StableLM 2 1.6B	210,26	68,17	142,09	217,77	71,59	146,18
StableLM 2 1.6B T	220,05	71,51	148,54	220,43	71,69	148,74
Gemma 2 2B	225,26	73,43	151,83	226,67	74,87	151,8
Gemma 2 2B T	221,88	74,51	147,37	223,6	76,37	147,24
Phi 3.5 mini 3.8B	225,68	52,82	172,86	233,35	59,13	174,22
Phi 3.5 mini 3.8B T	229,59	54,77	174,81	231,84	56	175,84
Qwen 3 4B	235,68	57,19	178,49	236	57,17	178,82
Qwen 3 4B T	233,24	57,25	175,99	237,02	58,01	179,01
Mistral 7B	216,4	56,36	160,04	236,5	63,64	172,86
Mistral 7B T	229,89	59,67	170,22	225,27	55,05	170,22
Qwen 2.5 7B	229,9	57,97	171,93	230,63	60,05	170,58
Qwen 2.5 7B T	192,36	54,46	137,89	233,74	60,13	173,61
LLaMA 3.1 8B	237,12	54,71	182,41	238,62	57	181,62
LLaMA 3.1 8B T	241,26	57,2	184,07	242,62	57,05	185,57
Gemma 3 12B	194,58	97,22	97,36	142,99	82,31	60,68
GPT OSS 20B	186,54	110,53	76,01	186,76	110,84	75,92
Gemma 3 27B	165,35	103,47	61,88	159,97	104,13	55,84
Qwen 3 30B	174,97	106,15	68,82	171,71	106	65,71
LLaMA 3.1 70B	213,78	118,64	95,14	188,93	109,78	79,15

Nota: QNO refere-se a Questão Não Original, enquanto QO refere-se a Questão Original.

A Figura 14 representa a potência média demandada em watts por modelo em questões não originais e a Figura 15 em questões originais.

5.1.4.3 Potência Máxima em watts

A Tabela 16 apresenta os resultados descritivos da potência máxima demandada (pico de potência instantânea) em watts obtidos pelos modelos avaliados.

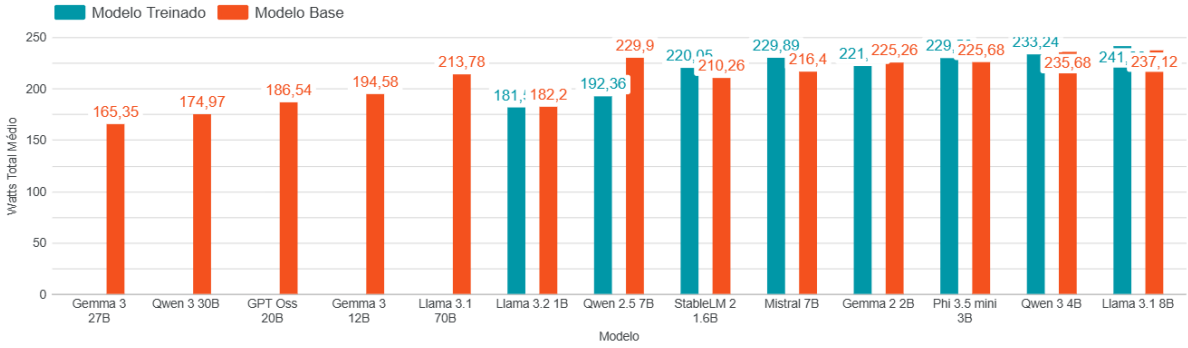


Figura 14 – Potência média demandada em watts por Modelo em Questões Não Originais

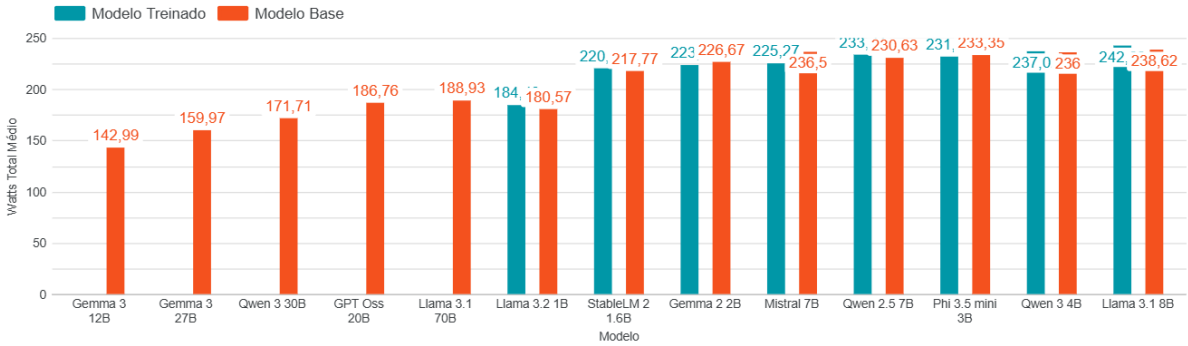


Figura 15 – Potência média demandada em watts por Modelo em Questões Originais

De maneira geral, observou-se que o máximo de potência tendeu a se aproximar dos limites físicos do hardware para a grande maioria dos modelos, resultando em uma distribuição de picos bastante alta e uniforme, que variou majoritariamente entre 260W e 340W. Os maiores picos globais de exigência foram registrados de forma compartilhada tanto por modelos grandes quanto médios, com destaque para *Gemma 3 27B*, *LLaMA 3.1 70B* e *Qwen 3 4B*, todos ultrapassando a faixa dos 300W de potência. Em contrapartida, as menores demandas máximas do estudo foram observadas no grande modelo *Qwen 3 30B*, que manteve seus picos controlados abaixo de 250W, demonstrando uma eficiência energética notável em momentos de estresse computacional.

Ao analisar o impacto direto do treinamento, observou-se um padrão de elevação nos picos de potência, afetando principalmente os modelos pequenos. Nos modelos *LLaMA 3.2 1B*, *StableLM 2 1.6B* e *Gemma 2 2B*, as versões treinadas provocaram aumentos expressivos na potência máxima nas questões não originais, saltando, por exemplo, de 248,51W para 292,32W. Esse comportamento indica que o ajuste fino induziu essas redes a utilizarem de forma mais agressiva e simultânea os recursos lógicos do sistema. A família Qwen repetiu sua tendência de estabilidade e controle arquitetural: enquanto o *Qwen 3 4B* manteve picos semelhantes ou reduzidos no cenário original, o *Qwen 2.5 7B* consolidou sua alta eficiência mantendo os menores picos, abaixo de 277W em todas as variações.

O tipo de questão exerceu um papel seletivo no comportamento dos picos de potência. Para as versões não treinadas dos modelos pequenos, as questões originais dispararam picos de watts substancialmente mais altos. Já para os modelos grandes, as questões originais geraram baixas oscilações nos picos totais, mas se concentraram nos dois componentes principais de hardware de maneira simultânea. Diferente da potência média, em que a CPU assumia o protagonismo isolado nos grandes modelos, a análise de picos máximos revela que o processamento atinge o teto operacional da GPU ao mesmo tempo em que exige o pico máximo da CPU, evidenciando que os cenários mais complexos extraem o limite absoluto do conjunto de hardware disponível.

Tabela 16 – Potência máxima em watts por modelo em questões

Modelo	Total QNO	CPU QNO	GPU QNO	Total QO	CPU QO	GPU QO
LLaMA 3.2 1B	248,51	90,29	158,22	299,39	136,88	162,51
LLaMA 3.2 1B T	292,32	142,3	150,02	293,46	135,61	157,85
StableLM 2 1.6B	274,53	97,42	177,11	323,54	145,37	178,21
StableLM 2 1.6B T	315,29	145,39	169,91	329,85	145,35	184,5
Gemma 2 2B	272,91	87,29	185,62	276,13	91,2	184,92
Gemma 2 2B T	318,08	145,29	172,8	333,01	145,33	187,68
Phi 3.5 mini 3.8B	273,75	83,74	190,01	286,41	97,56	188,85
Phi 3.5 mini 3.8B T	282,5	95,41	187,08	284,23	94,36	189,87
Qwen 3 4B	305,06	115,54	189,52	280	90,74	189,26
Qwen 3 4B T	335,28	139,91	195,37	284,69	95,66	189,03
Mistral 7B	285,89	93,09	192,8	279,98	88,07	191,9
Mistral 7B T	285,69	94,14	191,56	292,06	99,9	192,16
Qwen 2.5 7B	271,65	79,48	192,17	281,54	91,63	189,9
Qwen 2.5 7B T	264,03	75,76	188,28	276,48	85,01	191,47
LLaMA 3.1 8B	264,5	75,1	189,4	276,6	88,52	188,07
LLaMA 3.1 8B T	284,11	94,59	189,53	283,05	93,31	189,74
Gemma 3 12B	326,67	131,85	194,82	316,45	122,77	193,68
GPT OSS 20B	271,13	115,61	155,52	276,5	120,03	156,47
Gemma 3 27B	337,13	144,95	192,18	339,15	145,18	193,97
Qwen 3 30B	248,88	132,57	116,31	239,5	118,41	121,09
LLaMA 3.1 70B	336,48	145,42	191,06	338,37	145,37	193

Nota: QNO refere-se a Questão Não Original, enquanto QO refere-se a Questão Original.

A Figura 16 representa a potência máxima demandada em watts por modelo em questões não originais e a Figura 17 em questões originais.

5.2 Testes Estatísticos

Nesta subseção, são apresentados e discutidos os resultados dos testes estatísticos aplicados na avaliação empírica desta pesquisa. A análise está estruturada em três dimensões principais, de acordo com o plano metodológico: desempenho por complexidade, eficiência computacional e energética, e a dinâmica de *trade-offs* induzida pela adequação contextual e pelo ajuste fino.

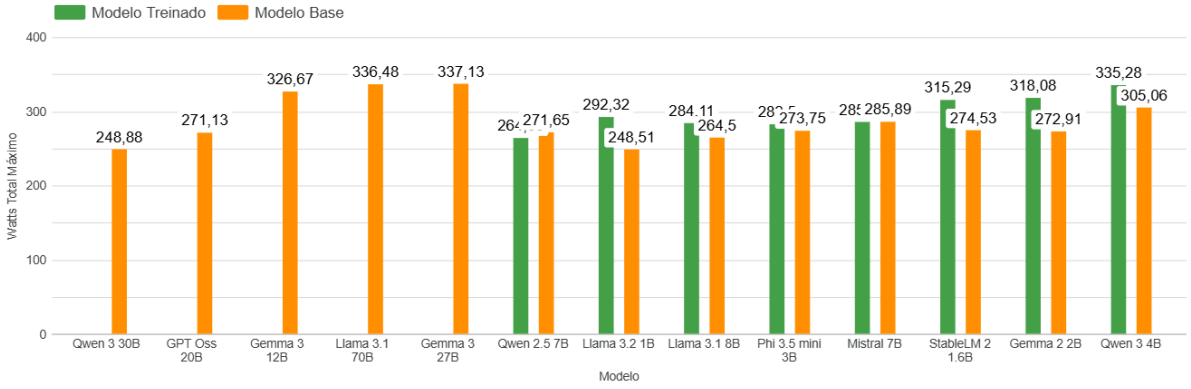


Figura 16 – Potência máxima demandada em watts por Modelo em Questões Não Originais

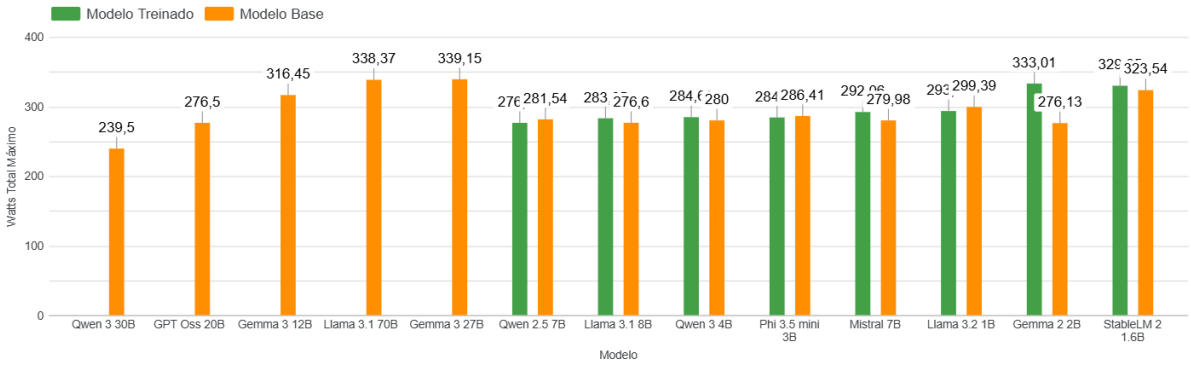


Figura 17 – Potência máxima demandada em watts por Modelo em Questões Originais

5.2.1 Dimensão 1: Acurácia por Complexidade

A primeira dimensão investiga a capacidade de acurácia dos modelos sob diferentes níveis de exigência cognitiva, respondendo diretamente às hipóteses H1 e H2.

A hipótese **H1** postulava que os SLMs de porte médio apresentariam desempenho equivalente aos LLMs em questões de nível Fácil e Médio. Contudo, os resultados do teste de equivalência TOST, configurado com margem de equivalência de $\Delta = 0{,}05$, não forneceram evidências de equivalência entre os modelos. Nas questões de nível Fácil, os LLMs obtiveram acurácia média de 84,86%, enquanto os SLMs alcançaram 72,29%, uma diferença de 12,57 pontos percentuais; o intervalo de confiança de 90% permaneceu inteiramente fora da margem estabelecida, indicando ausência de equivalência ($p_{TOST} = 0{,}993$). Esse mesmo padrão foi observado nas questões de nível Médio, nas quais os LLMs atingiram 72,58% de acerto, contra 56,13% dos SLMs, gerando uma diferença de 16,45 pontos percentuais, também sem evidência de equivalência ($p_{TOST} > 0{,}999$). De forma complementar, a regressão binomial via Equações de Estimação Generalizadas (GEE) confirmou essa discrepância, demonstrando que o uso de SLMs de porte médio reduz significativamente a probabilidade de acerto em relação à categoria de referência LLM ($\beta = -0{,}731$, $p = 0{,}001$), o que corresponde a uma redução nas chances de acerto de aproximadamente 51,9%.

Por outro lado, a hipótese **H2**, que postulava que LLMs apresentariam desempenho superior em tarefas de nível Difícil, foi plenamente suportada. O modelo GEE, utilizando contraste direcional, confirmou a vantagem estatisticamente significativa dos LLMs nesse cenário de maior complexidade cognitiva. Nessas tarefas, nas quais é demandada maior capacidade de generalização e raciocínio, os LLMs sustentaram uma acurácia de 56,95%, enquanto a acurácia dos SLMs decaiu para 36,68%, evidenciando uma diferença de aproximadamente 20 pontos percentuais. A análise da razão de chances (*Odds Ratio*) indicou que a probabilidade de um LLM acertar uma questão difícil foi 2,28 vezes superior à observada para os SLMs ($p < 0{,}001$, IC 95% [1,48; 3,52]). Tais evidências indicam que, em tarefas que exigem maior abstração e integração de conhecimento, as limitações paramétricas dos SLMs ainda representam um fator restritivo relevante, comprometendo seu desempenho em configurações *zero-shot* quando comparados aos modelos massivos.

5.2.2 Dimensão 2: Eficiência Computacional e Energética

A segunda dimensão investiga a latência, o custo computacional e o consumo energético entre SLMs e LLMs, respondendo diretamente às hipóteses de H3 a H7.

A hipótese **H3**, que postulava menor tempo de resposta dos SLMs em comparação aos LLMs, foi confirmada pelos resultados. Modelando-se a variável de tempo gasto em escala logarítmica, observou-se que a latência de geração dos LLMs foi, em média, 8,56 vezes superior à dos SLMs ($p < 0{,}001$; IC 95% [3,18; 23,08]). Esse efeito permanece estatisticamente significativo mesmo após o controle da estrutura hierárquica dos dados nos LMM, indicando que, em termos práticos, os SLMs produzem respostas substancialmente mais rápidas. A estimativa associada ao tipo de modelo reforça essa diferença, com coeficiente positivo e altamente significativo para LLMs em relação à categoria de referência SLM ($\beta = 2{,}148$, $p < 0{,}001$).

A hipótese **H4**, que postulava menor consumo de tokens pelos SLMs em relação aos LLMs, não foi suportada. A razão observada entre LLMs e SLMs foi de 0,91, com intervalo de confiança de 95% abrangendo a unidade [0,68; 1,22], o que indica ausência de diferença estatisticamente significativa no volume de tokens empregados na geração das respostas ($p = 0{,}531$). Isso evidencia que, embora a estimativa pontual sugira uma leve redução no consumo de tokens pelos LLMs, tal diferença não se mostra robusta o suficiente para sustentar a hipótese formulada. Com isso, os resultados sugerem que o custo em emissão de tokens permanece, na prática, comparável entre os dois grupos de modelos.

De forma análoga a H3, a hipótese **H5**, que postulava menor consumo energético dos SLMs em comparação aos LLMs, foi plenamente suportada. A análise por regressão OLS com erros-padrão robustos por cluster, corroborada pelo teste de Mann-Whitney, revelou que os LLMs consomem, em média, 4,83 vezes mais energia total do que os

modelos de menor porte ($p = 0{,}0003$ na regressão; $p < 0{,}001$ no teste não paramétrico; IC 95% [2,04; 11,45]). Esses achados demonstram uma desvantagem ambiental relevante dos LLMs, indicando que o ganho de desempenho em tarefas complexas é obtido às custas de um consumo energético significativamente superior, reforçando as vantagens de sustentabilidade preconizadas pela *Green AI* para os SLMs.

Em relação às hipóteses **H6** e **H7**, investigou-se se o custo computacional e o custo energético escalariam de maneira não linear conforme o aumento do porte dos modelos. Para isso, foram comparados modelos lineares e quadráticos por meio do Teste de Razão de Verossimilhança (LRT), utilizando modelos mistos com intercepto aleatório por arquitetura. No caso do custo computacional (**H6**), mensurado pelo número de tokens utilizados, não foram encontradas evidências de comportamento não linear em função da quantidade de parâmetros ($\chi^2 = 0{,}139$, $p = 0{,}709$). O termo quadrático associado ao porte do modelo também não apresentou significância estatística ($\beta = 0{,}021$, $p = 0{,}708$), indicando que o consumo de tokens permanece aproximadamente estável e linear ao longo da transição entre SLMs e LLMs, como pode ser observado na Figura 18. Esses resultados sugerem que modelos maiores não demandam quantidades desproporcionalmente superiores de tokens para produzir respostas.

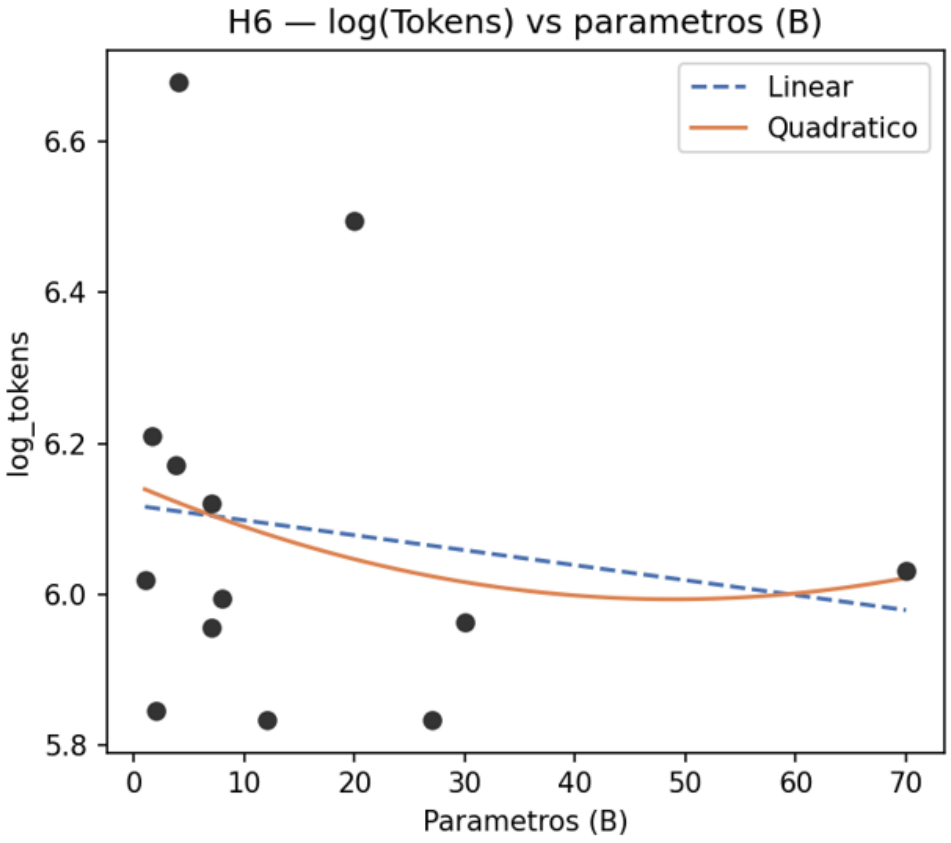


Figura 18 – log(Tokens) vs parâmetros (B)

De forma semelhante, a hipótese **H7**, que postulava um crescimento não linear

do consumo energético com o aumento do porte dos modelos, também não foi suportada pelos dados. A comparação entre os modelos linear e quadrático não revelou melhora estatisticamente significativa ao incorporar o termo de segunda ordem ($\chi^2 = 1{,}246$, $p = 0{,}264$). Além disso, o coeficiente quadrático estimado não apresentou significância estatística ($\beta = -0{,}181$, $p = 0{,}253$), indicando que, no espaço logarítmico analisado, o consumo energético cresce de maneira aproximadamente linear em relação ao número de parâmetros, conforme exibido na Figura 19. Em termos práticos, embora os LLMs consumam substancialmente mais energia, os resultados não indicam a existência de uma explosão exponencial de custo energético associada ao aumento de escala dentro do intervalo avaliado.

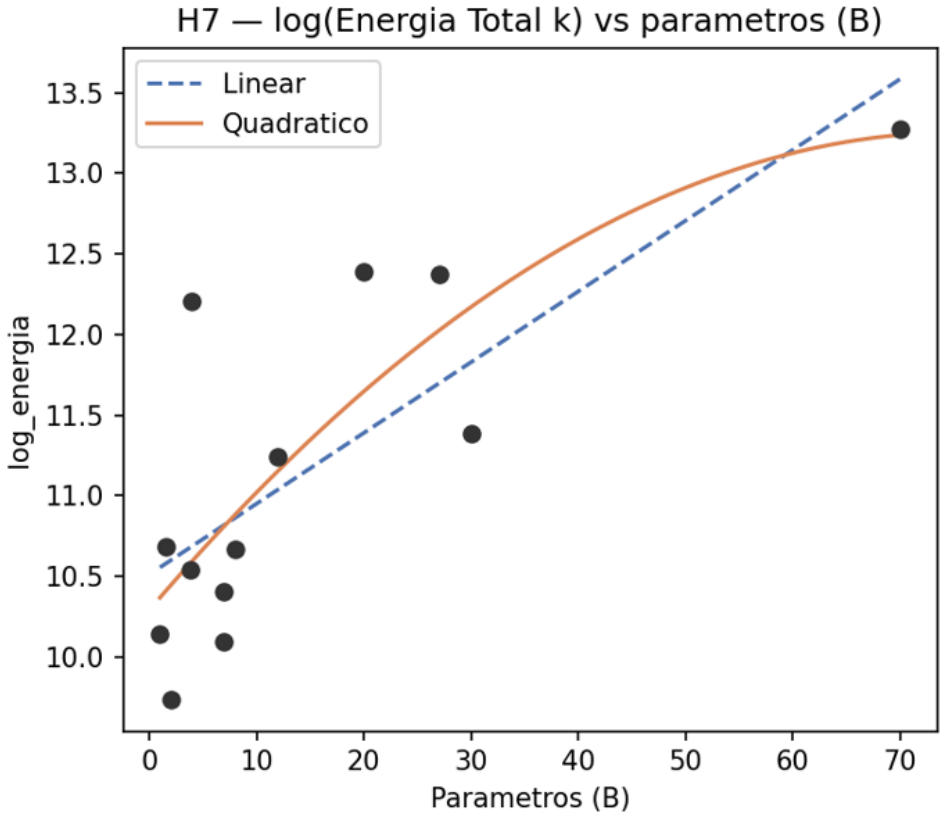


Figura 19 – log(Energia Total J) vs parâmetros (B)

5.2.3 Dimensão 3: *Trade-offs* e Adequação Contextual

A terceira dimensão consolida o conceito do *framework* ao analisar as interações de custo-benefício, respondendo diretamente às hipóteses H8 e H9.

A hipótese **H8**, que postulava a existência de *trade-offs* significativos entre acurácia e custo computacional, modulados por disciplina, nível de dificuldade e originalidade da questão, foi confirmada apenas de forma parcial, evidenciando que a relação entre esses dois fatores depende fortemente do contexto da tarefa. O efeito principal de volume de saída

log tokens foi negativo e estatisticamente significativo ($\beta = -0{,}935$; $p = 0{,}004$), indicando que, de modo geral, respostas mais longas estiveram associadas a menor probabilidade de acerto. Entretanto, esse padrão não se manteve uniforme em todas as condições analisadas. Entre as interações testadas, duas mostraram significância estatística: para questões de nível **Médio**, o aumento de tokens apresentou efeito compensatório positivo sobre a acurácia ($\beta = +0{,}508$; $p = 0{,}015$), sugerindo que, nesse nível de dificuldade, uma maior elaboração da resposta pode refletir maior esforço de raciocínio e, potencialmente, melhor desempenho; já na disciplina **Inglês**, o acréscimo de tokens exerceu efeito contrário, penalizando significativamente a acurácia ($\beta = -0{,}249$; $p = 0{,}024$). Em contrapartida, não foram observadas interações estatisticamente significativas para a originalidade da questão, o que indica ausência de modulação robusta desse fator sobre o efeito do comprimento da resposta. Assim, os resultados apontam que o custo computacional associado à verbosidade não é uniformemente benéfico nem prejudicial, mas varia conforme o tipo de conteúdo e o nível de complexidade da questão.

A hipótese **H9**, por sua vez, investigou se modelos treinados apresentariam ganho de acurácia em relação aos não treinados, considerando o impacto no custo computacional e energético. Os resultados mostraram que não houve ganho estatisticamente significativo de acurácia global para os modelos treinados em comparação à linha de base (OR = 1,04; $p = 0{,}722$), indicando que o *fine-tuning*, nas condições avaliadas, não produziu melhora cognitiva mensurável generalizada na tarefa. Em contrapartida, os custos operacionais cresceram de forma expressiva: as respostas tornaram-se, em média, 42% mais longas (razão = 1,42; $p < 0{,}001$), a latência quase quadruplicou (3,90×; $p < 0{,}001$) e o consumo energético aumentou em proporção geométrica de 3,08× ($p = 0{,}001$).

Esses achados revelam um severo descompasso entre desempenho e eficiência, no qual o ajuste fino, nas condições avaliadas, elevou drasticamente os custos de inferência sem entregar um benefício consistente de acurácia. Esse comportamento indica que técnicas de ajuste fino local sem rigorosas restrições arquiteturais para inferência podem comprometer o desempenho computacional e energético de um modelo, sem oferecer garantias de ganho cognitivo na resolução de questões educacionais. Tais intervenções podem, inclusive, induzir fenômenos de sobreajuste (*overfitting*) e contrariar os princípios da *Green AI*.

A análise integrada desses dados demonstra que a escolha por modelos não-treinados (*zero-shot*) baseados em conhecimento prévio, ou a seleção estratégica de SLMs para tarefas de baixa complexidade, constitui a via mais pragmática e sustentável para a democratização da inteligência artificial na educação.

5.3 Demonstração do Framework

Nesta subseção, são apresentados os resultados da demonstração do *framework* proposto com análise fundamentada na simulação de diferentes cenários educacionais, sendo estruturados a partir da Tabela 5. O objetivo principal desta etapa é demonstrar a capacidade de adaptação do algoritmo de *Profile Matching* na seleção do melhor modelo dentre os 21 disponíveis (13 base e 8 treinados) para cada contexto de uso, fornecendo o substrato analítico necessário para a discussão e validação das hipóteses H10 e H11.

5.3.1 Cenário 1: Tutoria em Tempo Real

O primeiro cenário avaliado no experimento simula um ambiente de tutoria por texto em que o aluno espera respostas imediatas. Neste contexto, a fluidez da conversa e o baixo custo de operação são prioridades, enquanto a complexidade profunda da resposta pode ser flexibilizada. O perfil-alvo (P_c) exigido foi parametrizado com acurácia e custo energético média e latência e custo computacional baixo.

Ao processar o vetor de requisitos em relação à base empírica contendo todos os modelos, o *framework* selecionou o modelo *Mistral 7B* em sua versão não treinada como a melhor opção, apresentando a menor distância euclidiana ($\Delta = 0,206$).

A seleção do *Mistral 7B* confirma perfeitamente com as análises descritivas das seções anteriores. Através da decomposição da distância multivariada, o modelo atingiu alinhamento exato (desvio quadrático zero) com os vetores ideais de acurácia e latência exigidos pelo cenário. Em comparação com os resultados anteriores, o modelo possui acurácia média de 53% (entre questões originais e não originais), situando-se na faixa "Média"de acurácia, além da sua latência (1,15s) estar perfeitamente atrelada à faixa "Baixa", garantindo o dinamismo necessário para uma tutoria via chat sem travar a interface do aluno.

As pequenas distâncias observadas no custo computacional e no consumo energético decorrem de o modelo demandar em média 478 tokens e 275 J por requisição, valores que ultrapassam de forma muito sutil os limites ótimos desejados para essas restrições, mas que, na apuração geral da função de minimização, representam o *trade-off* mais vantajoso do estudo.

A análise desse cenário fornece a primeira evidência empírica robusta em favor do framework multicritério, ao demonstrar que estratégias de seleção baseadas em métricas isoladas podem conduzir à escolha de modelos inadequados para aplicações educacionais em tempo real. A maximização exclusiva da acurácia resultaria na seleção do Qwen 3 30B, cuja elevada latência inviabilizaria interações instantâneas; por outro lado, a priorização apenas do menor custo favoreceria o Gemma 3 12B, comprometendo significativamente a eficiência energética. De modo semelhante, critérios focados exclusivamente em baixa

latência ou menor consumo energético conduziriam ao Gemma 2 2B, modelo que, apesar de sua fluidez operacional, apresentou desempenho crítico nas avaliações experimentais, tornando inviável a oferta de suporte pedagógico consistente.

Nesse contexto, o Mistral 7B, embora não lidere nenhuma métrica individualmente, revelou o maior equilíbrio global entre desempenho, custo e eficiência conforme as métricas esperadas, evidenciando matematicamente a importância de considerar os trade-offs inerentes à implantação de sistemas de inteligência artificial em ambientes educacionais reais.

5.3.2 Cenário 2: Correção em Lote (*Batch*)

O segundo cenário avaliado no experimento simula um ambiente de correção de avaliações em lote, como um ambiente em que há processamento assíncrono, na qual as respostas e atividades são processadas em segundo plano sem a expectativa de retorno imediato para o aluno. Neste contexto, a tolerância ao tempo de espera permite priorizar a qualidade da correção e a profundidade analítica, buscando um uso moderado da infraestrutura para não sobrecarregar os servidores. O perfil-alvo (P_c) exigido foi parametrizado com acurácia alta, latência alta e custo computacional e energético na faixa média.

Ao processar o vetor de requisitos em relação à base empírica contendo todos os modelos, o *framework* selecionou o modelo *GPT OSS 20B* em sua versão não treinada como a melhor opção, apresentando a menor distância euclidiana ($\Delta = 0,414$).

A seleção do *GPT OSS 20B* confirma perfeitamente as análises descritivas das seções anteriores. Através da decomposição da distância multivariada, o modelo demonstrou forte alinhamento com a exigência primária do cenário (desvio quadrático de apenas $0,0021$ na dimensão de acurácia). Em comparação com os resultados anteriores, o modelo possui uma excelente acurácia média de 73% (entre questões originais e não originais), situando-se confortavelmente na faixa "Alta". Além disso, sua latência de 10,82s se enquadra de forma coerente na tolerância de tempo exigida pela faixa "Alta", garantindo um processamento cadenciado e compatível com filas de execução assíncronas.

As distâncias maiores observadas no custo computacional e no consumo energético decorrem de o modelo demandar em média 690 tokens e 2.008 J por requisição, valores que ultrapassam sutilmente os limites ótimos desejados para essas restrições (faixa Média), mas que, na apuração geral da função de minimização, representam o *trade-off* mais vantajoso do estudo.

A análise desse cenário fornece mais uma evidência empírica robusta em favor do *framework* multicritério, ao demonstrar que estratégias de seleção baseadas em métricas isoladas podem conduzir à escolha de modelos inadequados para a dinâmica de processamento

proposta. A maximização exclusiva da acurácia resultaria na seleção do *Qwen 3 30B*, que não se adaptou tão bem ao vetor global de utilidade; por outro lado, a priorização apenas do menor custo de tokens favoreceria o *Gemma 3 12B*, que apresenta comportamentos distintos de consumo. De modo semelhante, critérios focados exclusivamente em baixa latência ou menor consumo energético conduziriam ao *Gemma 2 2B*, modelo que, apesar de sua leveza, apresentou desempenho crítico nas avaliações experimentais, tornando inviável a sua aplicação em correções pedagógicas que exigem alta confiabilidade.

Nesse contexto, o *GPT OSS 20B*, embora não lidere nenhuma métrica individualmente, revelou o maior equilíbrio global entre desempenho, custo e perfil de tempo para execuções em segundo plano, evidenciando matematicamente a importância de considerar os *trade-offs* inerentes à implantação de sistemas de inteligência artificial em ambientes educacionais reais.

5.3.3 Cenário 3: Monitoria Escalável para Grande Volume de Alunos

O terceiro cenário avaliado no experimento simula um ambiente de monitoria projetado para atender simultaneamente a um grande volume de alunos. Neste contexto, a capacidade de lidar com requisições escaláveis sem onerar excessivamente a geração de texto é fundamental para evitar gargalos na interface, aceitando que a infraestrutura em nuvem irá operar com uma tolerância maior ao consumo de energia para sustentar o volume. O perfil-alvo (P_c) exigido foi parametrizado com acurácia e custo computacional na faixa média, latência baixa e consumo energético alto.

Ao processar o vetor de requisitos em relação à base empírica contendo todos os modelos, o *framework* selecionou o modelo *Mistral 7B* em sua versão treinada como a melhor opção, apresentando a menor distância euclidiana ($\Delta = 0,506$).

A seleção do *Mistral 7B* treinado ilustra a capacidade do algoritmo de encontrar o modelo de maior aderência mesmo quando as restrições geram vetores de otimização conflitantes. Através da decomposição da distância multivariada, o modelo demonstrou um ajuste notável no custo computacional (desvio quadrático de apenas $0,0137$), consumindo em média 609 tokens por requisição, o que o alinha perfeitamente à faixa "Média"estipulada. Em relação à acurácia, o modelo obteve 41% de acertos, situando-se no limiar inferior da faixa "Média", o que atende ao requisito pedagógico básico do cenário de monitoria sem exceder a complexidade necessária.

As distâncias maiores observadas nesta seleção concentram-se no consumo energético e na latência, expondo o *trade-off* calculado pelo algoritmo. Embora o cenário tolerasse um gasto energético elevado (faixa Alta), o modelo demandou 1.186 J, gerando um afastamento matemático (desvio quadrático de $0,1786$) que, na prática, reflete uma folga operacional positiva. Por outro lado, o modelo apresentou um tempo de resposta de 5,25s, valor que

não atinge a fluidez extrema da faixa "Baixa"exigida, mas que, na apuração geral da função de minimização, representou o sacrifício mais vantajoso para manter a coerência das demais métricas. Modelos concorrentes mais rápidos, como as versões não treinadas do próprio *Mistral 7B* ou do *Phi 3.5 mini 3.8B*, foram preteridos por se distanciarem em outras dimensões críticas do vetor-alvo.

A análise desse cenário fornece mais uma evidência empírica robusta em favor do *framework* multicritério, ao demonstrar que estratégias de seleção baseadas em métricas isoladas podem conduzir à escolha de modelos inadequados para aplicações em larga escala. A maximização exclusiva da acurácia resultaria na seleção do *Qwen 3 30B*, cuja infraestrutura pesada inviabilizaria a escalabilidade ágil para milhares de alunos simultâneos; por outro lado, a priorização apenas do menor custo de tokens favoreceria o *Gemma 3 12B*, que não obteve o mesmo balanço estrutural. De modo semelhante, critérios focados exclusivamente em baixa latência ou menor consumo energético conduziriam ao *Gemma 2 2B*, modelo que, apesar de veloz, apresentou acurácia crítica nas avaliações experimentais, comprometendo de forma inaceitável a precisão do conteúdo da monitoria.

Nesse contexto, o *Mistral 7B* em sua versão treinada, embora não lidere nenhuma métrica individualmente, revelou o maior equilíbrio global para operar sob estresse de volume, evidenciando matematicamente a importância de considerar os *trade-offs* inerentes à implantação de sistemas de inteligência artificial em ambientes educacionais reais.

5.3.4 Cenário 4: Raciocínio Lógico Complexo (Sem Restrição de Hardware)

Direcionado a tarefas de alta exigência cognitiva, o quarto cenário explora a resolução de problemas complexos em um ambiente desprovido de restrições de infraestrutura computacional. Nesse cenário, é priorizada a máxima confiabilidade analítica (acurácia na faixa Alta) e aceita que o modelo seja mais verboso e detalhista (custo computacional na faixa Alta) e demande elevada carga elétrica (consumo energético Alto), exigindo apenas que a execução mantenha um tempo de resposta intermediário (latência Média) para preservar a interatividade do sistema.

A execução do *framework* multicritério através dessas diretrizes indicou o *Qwen 3 4B* em sua versão não treinada como a alternativa mais aderente, registrando a menor distância euclidiana do conjunto ($\Delta = 0,358$).

A seleção deste modelo confirma apontamento feitos na etapa de análise descritiva, na qual essa estrutura de menor porte superou diversos concorrentes maiores. Pela decomposição da métrica, nota-se que o *Qwen 3 4B* obteve um perfeito alinhamento nas faixas mais críticas (desvio quadrático de apenas $0,0021$ em acurácia e $0,0096$ em latência). Com 68% de precisão média geral, o modelo consolida sua posição na faixa "Alta"de acurácia, entregando as respostas em 7,15s, o que se encaixa rigorosamente na

tolerância estipulada pela faixa "Média"de tempo.

Em relação aos custos operacionais, as métricas capturadas refletem o comportamento esperado para a formulação de cadeias de pensamento mais densas. A aceitação de um maior esforço pelo cenário (vetores de custo e energia em faixa "Alta") justifica os 812 tokens gerados e os 1.665 J exigidos por requisição. O balanço foi muito favorável, até que o segundo colocado no ranqueamento foi a própria versão treinada do modelo (*Qwen 3 4B T*, com $\Delta = 0,375$), evidenciando a vocação dessa família arquitetural para tarefas que exigem raciocínio extenso.

Sob a ótica da validação metodológica, este cenário reforça mais uma vez a fragilidade de decisões baseadas em métricas isoladas. Caso a busca priorizasse estritamente a maximização da acurácia, o modelo selecionado seria o grande *Qwen 3 30B*. No entanto, essa escolha tradicional falharia em atender ao perfil completo desejado: o *Qwen 3 30B* consome em média apenas 408 tokens (faixa Baixa), demonstrando uma penalização em frente a um cenário que exige uma alta explicabilidade e detalhamento (custo computacional Alto). Em outro ponto, o foco na agilidade ou na economia de energia apontaria para o *Gemma 2 2B*, cuja baixa capacidade cognitiva inviabilizaria qualquer tentativa de resolução lógica.

Portanto, o *Qwen 3 4B* não apenas aparece como vencedor de um ranqueamento unidimensional, mas sim como o encaixe ideal entre uma inferência altamente capaz e detalhista em um ambiente que está disposto a pagar o preço energético e temporal para obter esse nível de raciocínio lógico.

5.3.5 Cenário 5: Quiz Gamificado Mobile (Leve e Ágil)

O próximo cenário avaliado simula um aplicativo de quiz gamificado para dispositivos móveis. Nesse contexto, a prioridade é oferecer uma experiência rápida e que consuma o mínimo possível de bateria e dados do celular do usuário. A exigência por respostas altamente complexas e precisas é menor, já que o foco é a agilidade e o engajamento do jogo. Por isso, o perfil-alvo (P_c) foi configurado de forma a exigir o menor esforço possível da máquina: níveis baixos para acurácia, latência, custo computacional e consumo energético.

Ao processar essas exigências na base de dados, o algoritmo selecionou o modelo *Gemma 2 2B* em sua versão não treinada como a opção ideal. Ele apresentou a menor distância euclidiana registrada no estudo até o momento, com um ajuste quase perfeito em relação ao que foi pedido ($\Delta = 0,098$).

Essa escolha está totalmente alinhada com os resultados descritivos vistos nas seções anteriores. A decomposição da distância mostra que o *Gemma 2 2B* se encaixou de forma excelente em todas as dimensões, com desvios muito próximos de zero. Sua acurácia

média de 23% o coloca exatamente na faixa "Baixa"solicitada. Além disso, o modelo se destacou por uma velocidade impressionante, com um tempo de resposta de apenas 0,61s, garantindo a fluidez imediata necessária para um jogo de celular sem travamentos.

O modelo também atendeu perfeitamente às restrições de infraestrutura do cenário. Ele consumiu em média apenas 370 tokens por requisição e demandou o menor gasto de energia de todo o experimento (141 J). O segundo colocado do ranqueamento, o *LLaMA 3.2 1B* em sua versão base ($\Delta = 0,238$), ficou para trás porque, apesar de ser ainda menor em tamanho de parâmetros, apresentou o dobro do tempo de resposta (1,22s) e consumiu mais energia, o que o afastou do equilíbrio ideal calculado pelo *framework*.

A análise deste cenário traz uma observação diferente e importante sobre a escolha por métricas isoladas. Ao contrário dos cenários anteriores, o modelo vencedor (*Gemma 2 2B*) também seria o escolhido se o desenvolvedor do aplicativo olhasse apenas para a "menor latência"ou para a "menor energia". No entanto, o uso do *framework* multicritério se mostra essencial porque protege o sistema de decisões desastrosas.

Em um cenário de escolha baseada apenas no topo de acurácia, o *Qwen 3 30B* seria o selecionado, o que travaria o celular do usuário devido ao seu peso computacional. Se o foco estivesse apenas em utilizar o mínimo de tokens possível, o *Gemma 3 12B* venceria, mas gastaria muito mais bateria do dispositivo. Dessa forma, o *framework* prova matematicamente que o *Gemma 2 2B* não foi selecionado apenas por ser rápido, mas por ser a única estrutura capaz de combinar rapidez, respostas curtas e baixo consumo de energia simultaneamente.

5.3.6 Cenário 6: Simulado ENEM com Feedback Explicativo

O próximo cenário avaliado simula uma plataforma de simulados para o Exame Nacional do Ensino Médio (ENEM) com foco na geração de feedbacks explicativos detalhados. Nesse contexto, a precisão da resposta é fundamental para garantir a qualidade pedagógica (acurácia na faixa Alta) e espera-se que o modelo utilize uma quantidade moderada de tokens para explicar detalhadamente os erros e acertos (custo computacional na faixa Média). Como as questões são respondidas individualmente pelo estudante, uma espera intermediária é aceitável (latência Média), mas a infraestrutura precisa tolerar picos de consumo de energia gerados pela alta demanda típica desse tipo de exame (consumo energético Alto).

Ao processar essas restrições na base de dados, o *framework* selecionou o modelo *GPT OSS 20B* em sua versão não treinada como a opção ideal, apresentando a menor distância euclidiana com valor de $\Delta = 0,461$.

A escolha do *GPT OSS 20B* está em perfeito alinhamento com os resultados descritivos das seções anteriores. Na decomposição da distância multivariada, o modelo

mostrou um ajuste quase exato no critério mais importante do cenário: a acurácia (desvio quadrático de apenas $0,0021$). O modelo obteve uma taxa de acurácia de 73%, posicionando-se de forma ideal dentro da faixa "Alta". Para a construção de respostas detalhadas, o modelo demandou em média 690 tokens por requisição, um valor sutilmente superior ao ponto médio da faixa "Média"(550 tokens), mas perfeitamente adequado para dar um retorno completo ao estudante.

Os maiores afastamentos ocorreram nas dimensões de latência e consumo de energia, refletindo os *trade-offs* calculados pelo algoritmo. O tempo de resposta foi de 10,82s, o que ultrapassou a meta da faixa "Média"(estabelecida em torno de 5s), mas permaneceu dentro de um limite tolerável para um simulado focado em estudo profundo. Quanto à energia, o gasto médio foi de 2.008 J. Embora esse valor seja inferior ao limite máximo da faixa "Alta", a folga foi computada pelo algoritmo para equilibrar a decisão global. Modelos concorrentes de grande porte, como o *Qwen 3 30B* ($\Delta = 0,598$), ficaram em desvantagem por entregarem textos curtos demais (408 tokens), afastando-se do perfil de feedback detalhado exigido.

A análise deste cenário reforça a importância de um sistema de decisão multicritério em oposição a escolhas baseadas em uma única métrica. Se apenas a maior acurácia fosse considerada, o modelo escolhido seria o *Qwen 3 30B*, o que deixaria o simulado com explicações superficiais e compactas devido ao seu baixo número de tokens gerados. Por outro lado, focar estritamente em reduzir custos com tokens levaria à escolha do *Gemma 3 12B*, que não atenderia ao perfil de consumo elétrico planejado. Da mesma forma, escolher apenas pela menor latência ou menor energia apontaria para o *Gemma 2 2B*, um modelo rápido, mas que falharia gravemente na tarefa pedagógica devido ao seu baixo índice de acurácia (23%).

Nesse contexto, o *GPT OSS 20B* provou ser a alternativa mais coerente. Mesmo sem liderar nenhuma métrica de forma isolada, ele foi o único modelo capaz de equilibrar uma alta taxa de acertos com um volume de texto adequado para a correção, evidenciando matematicamente a importância de considerar os *trade-offs* inerentes à implantação de sistemas de inteligência artificial em ambientes educacionais reais.

5.3.7 Cenário 7: Sala de Aula Premium Multimodal

O cenário analisado a seguir propõe o contexto de uma sala de aula de alto padrão tecnológico e com suporte multimodal. Nesse ambiente, busca-se entregar respostas rápidas e de alta qualidade pedagógica para os estudantes, partindo do princípio de que a infraestrutura local ou o orçamento de nuvem não possuem restrições. Assim, é tolerável um gasto elevado de processamento e energético para manter a excelência do serviço. A parametrização do perfil-alvo (P_c) seguiu definindo níveis altos para acurácia, custo computacional e consumo de energia, combinados com uma baixa latência.

Diante dessas condições, o *framework* apontou o *Qwen 3 4B* em sua versão treinada como a solução mais equilibrada, alcançando a menor distância euclidiana global ($\Delta = 0,433$).

Os resultados práticos confirmam os padrões observados nas análises anteriores para esse modelo. Na decomposição multivariada, o *Qwen 3 4B* treinado atingiu um alinhamento muito próximo no critério de acurácia (desvio quadrático de $0,0027$). O modelo obteve uma acurácia média de 67%, o que o consolida firmemente na faixa "Alta"de acertos esperada para dar suporte a uma aula de alto nível, garantindo respostas confiáveis aos alunos.

Em relação ao custo computacional, o modelo obteve um ajuste perfeito (desvio quadrático zero) com a meta, demandando em média 753 tokens por requisição, o que preenche o requisito de respostas detalhadas e completas (faixa Alta) exigido pelo contexto multimodal. O principal desafio técnico refletido na distância deu-se na latência, que registrou 6,18s — um tempo que se afasta do ideal imediato da faixa "Baixa", mas que se mostrou inevitável diante da carga de processamento. Quanto ao consumo energético, o gasto real foi de 1.446 J, porém embora o cenário estivesse aberto a consumos bem maiores (faixa Alta), essa folga foi assimilada pelo algoritmo no cálculo da utilidade final. O equilíbrio da família *Qwen* ficou evidente com o segundo lugar do ranking ocupado por sua versão base não treinada ($\Delta = 0,435$).

A avaliação desse caso reforça o risco de desenhar sistemas educacionais olhando para indicadores isolados. Uma abordagem focada puramente na maior acurácia escolheria o *Qwen 3 30B*, mas seu custo computacional (408 tokens) falharia em entregar o detalhamento demandado pelo perfil multimodal. Por outro lado, focar apenas na menor latência ou menor energia levaria à escolha do *Gemma 2 2B*, que traria uma velocidade excelente (0,6s), mas comprometeria drasticamente o conteúdo pedagógico devido à sua baixa taxa de acerto (23%). Reduzir custos de tokens isoladamente traria o *Gemma 3 12B*, desatendendo as demais variáveis de consumo.

Com isso, o *Qwen 3 4B* treinado se consolida como o modelo ideal para este cenário. Ele demonstra que, em ambientes com infraestrutura avançada, a melhor escolha não é necessariamente o maior ou o mais rápido modelo de forma isolada, mas sim aquele que resolve de maneira harmônica as tensões entre precisão pedagógica, extensão textual e tempo de resposta.

5.3.8 Cenário 8: Escola com Recursos Limitados e Conectividade Instável

O oitavo cenário avaliado simula a realidade de uma instituição de ensino com recursos computacionais limitados e conectividade de rede instável. Nesse ambiente, o sistema não pode sobrecarregar os servidores locais nem gerar alto tráfego de dados,

exigindo o mínimo de impacto na infraestrutura. Por essa razão, busca-se um patamar intermediário para a qualidade das respostas e o tempo de atendimento (acurácia e latência na faixa Média), enquanto a prioridade absoluta é garantir a máxima economia de tokens e eletricidade (custo computacional e consumo energético na faixa Baixa).

Ao processar esse perfil de restrições na base empírica, o *framework* selecionou o modelo *LLaMA 3.1 8B* em sua versão não treinada como a alternativa ideal, registrando uma distância euclidiana de $\Delta = 0,229$.

A escolha do *LLaMA 3.1 8B* base demonstra o equilíbrio do algoritmo diante de restrições rígidas de infraestrutura. Pela decomposição da distância multivariada, o modelo alcançou uma acurácia média de 42%, situando-se de forma adequada na faixa "Média"para dar um suporte pedagógico básico sem a necessidade de uma arquitetura gigante. No quesito velocidade, sua latência média de 1,50s superou com folga o centro projetado para a faixa "Média"(5,00s), garantindo respostas rápidas que ajudam a mitigar os problemas de uma conexão instável.

O grande destaque do modelo foi o seu encaixe quase perfeito nas dimensões de consumo, apresentando distâncias mínimas no custo de tokens $(0,0021)$ e na energia $(0,0004)$. O *LLaMA 3.1 8B* demandou em média 428 tokens e consumiu apenas 360 J por requisição, mantendo-se muito próximo dos limites ideais da faixa "Baixa". Essa otimização foi tão acirrada que o modelo superou o *Mistral 7B* em sua versão base (segundo colocado com a mesma distância aproximada de $\Delta = 0,229$) devido à consistência combinada dessas quatro dimensões.

O confronto com as escolhas baseadas em métrica única reforça o valor prático do *framework*. Se a escolha do modelo se baseasse apenas na ótica de maior acurácia, o modelo selecionado seria o *Qwen 3 30B*, cuja robustez inviabilizaria a execução em computadores escolares antigos ou conexões lentas. Por outro lado, focar estritamente em menor latência ou menor consumo elétrico apontaria para o *Gemma 2 2B*, que traria rapidez, mas comprometeria o aprendizado devido ao seu baixo índice de acurácia. Se o critério fosse apenas o menor custo de tokens, o *Gemma 3 12B* venceria, mas sem o mesmo balanço energético.

Nesse contexto, o *LLaMA 3.1 8B* base provou ser a solução mais viável. Sem liderar nenhum indicador isoladamente, ele se consolidou como o modelo ideal por proteger a infraestrutura limitada da escola, garantindo tráfego leve e baixo consumo elétrico sem abrir mão da velocidade e precisão necessárias para o ambiente educacional.

5.3.9 Cenário 9: Personalização Adaptativa de Trilhas de Aprendizagem

O nono cenário avaliado simula um sistema de recomendação contínua que adapta trilhas de aprendizagem conforme o progresso do aluno. Para que o sistema faça boas

recomendações pedagógicas, a exigência por precisão é elevada (acurácia na faixa Alta). No entanto, como essa adaptação ocorre em segundo plano a cada interação do estudante, a infraestrutura não necessita operar em limites altos, exigindo um comportamento equilibrado e sustentável nas demais métricas (latência, custo computacional e consumo energético na faixa Média).

Ao processar esse vetor de requisitos na base de dados, o *framework* selecionou o modelo *Phi 3.5 mini 3.8B* em sua versão não treinada como a solução mais adequada, registrando uma distância euclidiana de $\Delta = 0,281$.

A indicação deste modelo demonstra a capacidade do algoritmo de encontrar o melhor ajuste prático quando um modelo de menor porte entrega um rendimento excepcional. Observando a decomposição da distância, o modelo atingiu uma acurácia média de 60%, porém mesmo esse valor pertencendo ao topo da faixa "Média"(e não estritamente à faixa "Alta"solicitada), o afastamento matemático foi tolerado $(0,0267)$ porque o modelo compensou essa leve diferença com um alinhamento excelente nas dimensões de custo. O custo computacional médio de 508 tokens por requisição garantiu um encaixe quase perfeito com a faixa "Média"de custo computacional (distância de apenas $0,0070$).

Nas métricas de operação, o modelo entregou respostas em 1,37s e demandou apenas 316 J. Embora o cenário aceitasse patamares "Médios"para esses fatores, a operação nas faixas "Baixas"gerou um saldo computacional vantajoso na minimização da utilidade geral. Esse balanço harmônico fez com que o *Phi 3.5 mini 3.8B* superasse o poderoso *Qwen 3 30B* (segundo colocado com $\Delta = 0,301$), que apesar de entregar a acurácia "Alta"desejada (76%), distanciou-se do perfil ideal devido a um menor volume de tokens e características de consumo diferentes.

O confronto metodológico contra as escolhas baseadas em métrica única reafirma o valor do *Profile Matching*. Se o sistema de trilhas fosse parametrizado apenas para buscar a maior acurácia absoluta, o *Qwen 3 30B* seria o escolhido, entregando textos possivelmente mais curtos que o idealizado para a explicação do caminho adaptativo. Caso a prioridade fosse apenas a velocidade instantânea ou a maior economia de energia, o sistema apontaria para o *Gemma 2 2B*, cuja baixa taxa de acertos (23%) desorientaria a trilha de estudos do aluno.

Dessa forma, o *Phi 3.5 mini 3.8B* emerge como a decisão mais racional. Ele comprova que, em aplicações de adaptação contínua, o modelo ideal não precisa ser o líder absoluto de nenhum *ranking* isolado, mas sim aquele que garante boas decisões pedagógicas sem desestabilizar o orçamento computacional da plataforma.

5.3.10 Cenário 10: *Baseline* Equilibrado para AVA Geral

O décimo e último cenário propõe uma configuração equilibrada, servindo como uma linha de base para um Ambiente Virtual de Aprendizagem (AVA) genérico. O objetivo não é focar em nenhuma métrica extrema, mas sim encontrar o modelo mais equilibrado para atender a demandas cotidianas variadas. Por isso, o perfil-alvo (P_c) foi parametrizado com exigências na faixa "Média"para todos os quatro critérios: acurácia, latência, custo computacional e consumo energético.

Ao submeter essa configuração neutra ao *framework*, o algoritmo indicou o *Mistral 7B* em sua versão submetida ao ajuste fino (treinada) como o modelo de maior adequação, registrando a menor distância euclidiana da simulação ($\Delta = 0,218$).

Observando a decomposição da distância, o modelo treinado cravou um ajuste praticamente perfeito na latência (desvio quadrático de apenas $0,0001$), entregando respostas em 5,25s, valor esperado para a faixa "Média". O desempenho no consumo energético (1.186 J) e no custo computacional (609 tokens) também acompanhou de perto essa tendência central, garantindo que o modelo opere de maneira previsível, sem picos de lentidão ou de exigência na máquina. A maior distância na análise ocorreu na acurácia ($0,0311$), já que o modelo obteve 41% de acurácia. Embora esse valor fique no limite inferior da faixa "Média", o *framework* o aceitou como o sacrifício necessário para manter as outras três métricas perfeitamente alinhadas.

É interessante notar o comportamento dos concorrentes diretos neste cenário. Modelos não treinados, como o *Phi 3.5 mini 3.8B* (segundo colocado) e a própria versão base do *Mistral 7B* (terceiro colocado), apresentaram índices elevados de acurácia, porém por serem excessivamente rápidos e econômicos, eles acabaram se distanciando do centro geométrico "Médio"exigido para este AVA genérico, o que explica a vantagem da versão treinada.

Ao encontrar o *Mistral 7B* treinado como o ponto de equilíbrio empírico deste estudo, é demonstrado que a seleção de um modelo de linguagem para a educação não deve ser tratada como uma análise de métricas isoladas, mas sim como um exercício rigoroso de adequar as características de desempenho do modelo à realidade e às restrições do contexto de aplicação.

5.3.11 Síntese Geral e Hipóteses

A execução de diferentes cenários utilizando um *framework* multicritério por meio da abordagem de *Profile Matching* consolidou uma visão abrangente sobre o comportamento das arquiteturas avaliadas. Ao submeter a base de dados dos modelos aos dez cenários educacionais distintos, foi possível observar como as métricas de acurácia, latência, custo computacional e eficiência energética interagem de maneira sistêmica. Os resultados obtidos

nesta etapa finalizam a dimensão de avaliação da pesquisa *Design Science Research* e fornecem as evidências definitivas para a discussão das duas hipóteses centrais do estudo.

A Tabela 17 apresenta os modelos selecionados e as respectivas distâncias euclidianas para cada cenário avaliado.

Tabela 17 – Modelo selecionado e distância euclidiana por cenário

Cenário	**Modelo Escolhido**	**Tipo de Modelo**	**Distância Euclidiana**
Tutoria em Tempo Real	*Mistral 7B*	Modelo Base	0,206
Correção em Lote (*Batch*)	*GPT OSS 20B*	Modelo Base	0,414
Monitoria Escalável para Grande Volume de Alunos	*Mistral 7B*	Modelo Treinado	0,506
Raciocínio Lógico Complexo (Sem Restrição de Hardware)	*Qwen 3 4B*	Modelo Base	0,358
Quiz Gamificado Mobile (Leve e Ágil)	*Gemma 2 2B*	Modelo Base	0,098
Simulado ENEM com Feedback Explicativo	*GPT OSS 20B*	Modelo Base	0,461
Sala de Aula Premium Multimodal	*Qwen 3 4B*	Modelo Treinado	0,433
Escola com Recursos Limitados e Conectividade Instável	*LLaMA 3.1 8B*	Modelo Base	0,229
Personalização Adaptativa de Trilhas de Aprendizagem	*Phi 3.5 mini 3.8B*	Modelo Base	0,281
Baseline Equilibrado para AVA Geral	*Mistral 7B*	Modelo Treinado	0,218

Nota: Modelos classificados como *Modelo Base* correspondem às versões originais disponibilizadas pelos desenvolvedores. Modelos classificados como *Modelo Treinado* correspondem às versões submetidas ao processo de ajuste fino (*fine-tuning*) realizado neste estudo.

A **Hipótese H10** postula que não existe um modelo universalmente superior em todos os cenários operacionais e que a escolha ideal depende da combinação específica de exigências de cada ambiente. A análise global dos dez cenários confirma esta hipótese de forma categórica.

Ao longo das execuções, o *framework* não tendeu sua escolha para um único modelo, pelo contrário, em que oito modelos diferentes foram selecionados para atender aos dez perfis-alvo: *Mistral 7B* (Base e Treinado), *GPT OSS 20B* (Base), *Qwen 3 4B* (Base e Treinado), *Gemma 2 2B* (Base), *LLaMA 3.1 8B* (Base) e *Phi 3.5 mini 3.8B* (Base). Essa dispersão comprova que as forças e fraquezas de SLMs e LLMs são intrinsecamente contextuais.

LLMs, como o *GPT OSS 20B*, provaram-se indispensáveis em cenários mais profundos, no qual há tolerância de um maior consumo energético e o maior tempo

de latência, permitindo extrair uma alta acurácia. Em contrapartida, cenários com recursos limitados demandaram arquiteturas de altíssima eficiência energética e baixa latência, justificando a escolha de SLMs, como o *LLaMA 3.1 8B* e o *Gemma 2 2B*. Ao provar matematicamente que a adequação de um modelo muda radicalmente conforme o vetor de utilidade do cenário, a hipótese H10 é confirmada e referendada pelos dados experimentais.

A **Hipótese H11** afirma que a aplicação de um *framework* multicritério baseado em aproximação de perfis seleciona modelos mais adequados à realidade do usuário do que uma escolha baseada na maximização de métricas isoladas. Os testes de confronto validam integralmente esta premissa.

Se a seleção arquitetural desta pesquisa fosse guiada pela lógica tradicional dos *rankings* unidimensionais, que priorizam quase exclusivamente a acurácia, o modelo *Qwen 3 30B* teria sido imposto a todos os cenários por possuir a maior acurácia geral do estudo. Contudo, o *framework* revelou que a adoção deste modelo causaria falhas em alguns contextos, sobrecarregaria redes instáveis e geraria explicações curtas demais em cenários que exigiam alto volume de tokens para feedback pedagógico.

De maneira análoga, se o critério fosse pautado apenas pela vertente puramente sustentável (menor consumo energético) ou pela baixa latência, o modelo *Gemma 2 2B* seria o mais escolhido. Porém, essa aplicação destruiria a confiabilidade de sistemas educacionais devido à sua baixa taxa de acurácia em questões mais complexas.

O algoritmo multicritério evidenciou que a verdadeira inteligência na escolha de modelos não reside em encontrar o pico de uma única métrica, mas sim em gerenciar os *trade-offs*. Em 90% dos cenários, o modelo ótimo indicado pela distância euclidiana não era o líder de nenhuma métrica individual, mas sim a arquitetura que ofereceu a geometria computacional mais harmônica entre inteligência, fluidez, prolixidade e gasto energético. Portanto, a hipótese H11 é plenamente confirmada, demonstrando que o artefato proposto atende de forma superior às restrições multidimensionais que caracterizam a implantação de inteligência artificial no mundo real.

Em síntese, a execução dos dez cenários evidencia que o *framework* proposto não privilegia universalmente modelos maiores, menores, base ou treinados, mas seleciona a arquitetura cuja configuração melhor se ajusta ao perfil de utilidade de cada contexto. Esse resultado reforça que a escolha de modelos de linguagem para aplicações educacionais deve ser compreendida como um problema de decisão multicritério, no qual a superioridade em uma métrica isolada não implica adequação global ao cenário de uso, corroborando as hipóteses H10 e H11.

5.4 Considerações Finais

Esta subseção sintetiza, de forma integrada, os principais achados da pesquisa, articulando os resultados experimentais, os testes estatísticos e a demonstração do *framework* proposto.

Em relação à acurácia por complexidade, os resultados não confirmaram a hipótese **H1**, segundo a qual os SLMs de porte médio apresentariam desempenho equivalente aos LLMs em questões de níveis Fácil e Médio. Em ambos os níveis, os LLMs apresentaram acurácia superior, com diferenças de 12,57 e 16,45 pontos percentuais, respectivamente, sem evidências de equivalência estatística. Em contrapartida, a hipótese **H2** foi plenamente suportada, uma vez que os LLMs mantiveram desempenho significativamente superior nas questões de nível Difícil, com diferença de aproximadamente 20 pontos percentuais e probabilidade de acerto cerca de 2,28 vezes maior. Esses achados indicam que, embora os SLMs de porte médio sejam competitivos em determinados contextos, suas limitações paramétricas ainda restringem o desempenho em tarefas de maior complexidade.

Quanto à eficiência computacional e energética, as evidências confirmaram as hipóteses **H3** e **H5**: os SLMs apresentaram latência substancialmente menor, gerando respostas, em média, 8,56 vezes mais rápidas que os LLMs, e também consumiram significativamente menos energia, ao passo que os LLMs demandaram cerca de 4,83 vezes mais energia. Esses resultados reforçam as vantagens de sustentabilidade associadas à *Green AI* e evidenciam que modelos de menor porte podem oferecer ganhos relevantes de eficiência operacional.

Em contrapartida, a hipótese **H4** não foi suportada, pois o consumo de tokens mostrou-se estatisticamente comparável entre SLMs e LLMs. Além disso, as hipóteses **H6** e **H7** também não foram confirmadas, uma vez que tanto o custo computacional quanto o consumo energético cresceram de maneira aproximadamente linear em relação ao número de parâmetros, sem evidência de explosão exponencial de custo associada ao aumento de escala.

A análise dos *trade-offs* produziu um dos achados mais relevantes do estudo, especialmente no que se refere ao efeito do ajuste fino. A hipótese **H8** foi confirmada apenas parcialmente: respostas mais longas estiveram, de modo geral, associadas a menor probabilidade de acerto, mas esse efeito variou conforme o contexto, mostrando-se compensatório em questões de nível Médio e penalizador na disciplina de Inglês.

A hipótese **H9**, por sua vez, não foi suportada. O ajuste fino, nas condições avaliadas, não produziu ganho de acurácia estatisticamente significativo, ao mesmo tempo em que elevou expressivamente os custos operacionais: as respostas tornaram-se, em média, 42% mais longas, a latência praticamente quadruplicou e o consumo energético cresceu em proporção de 3,08 vezes. Esse descompasso entre desempenho e eficiência sugere

que técnicas de ajuste fino local, sem restrições arquiteturais rigorosas, podem induzir sobreajuste (*overfitting*) e contrariar os princípios da *Green AI*, sem oferecer garantias de ganho cognitivo.

A demonstração do artefato consolidou as duas hipóteses centrais da pesquisa. A hipótese **H10** foi confirmada de forma robusta: ao longo dos dez cenários educacionais simulados, o *framework* não convergiu para um único modelo, mas selecionou oito modelos distintos, evidenciando que as forças e fraquezas de SLMs e LLMs são intrinsecamente contextuais.

Da mesma forma, a hipótese **H11** foi plenamente suportada, demonstrando que a seleção baseada em aproximação de perfis supera a escolha guiada pela maximização de métricas isoladas. Em 90% dos cenários, o modelo indicado como ideal não era o líder de nenhuma métrica individual, mas aquele que oferecia o equilíbrio mais adequado entre acurácia, latência, custo computacional e consumo energético para o cenário proposto. Em síntese, estratégias unidimensionais tenderiam a conduzir a escolhas inadequadas, seja pela elevada latência, seja pela perda de acurácia em contextos que exigem maior confiabilidade pedagógica.

De modo geral, os resultados desta pesquisa demonstram que a adequação entre modelo e contexto é o elemento central para a tomada de decisão em aplicações educacionais baseadas em IA. Assim, a principal contribuição do estudo está em evidenciar que a escolha entre SLMs e LLMs não deve ser guiada apenas por uma métrica individual ou porte do modelo, mas pela combinação entre desempenho, eficiência e pertinência ao cenário de uso.

6 Conclusão

Esta pesquisa teve como objetivo propor e validar um *framework* de decisão multicritério para a seleção entre *Small Language Models* (SLMs) e *Large Language Models* (LLMs) em contextos educacionais.

Atualmente, a literatura aponta uma lacuna de estudos sistemáticos que investiguem, de forma comparativa e contextualizada, o comportamento de SLMs e LLMs em diferentes cenários educacionais, considerando simultaneamente métricas de desempenho, custo e sustentabilidade. Essa necessidade torna-se ainda mais urgente diante das crescentes preocupações associadas ao impacto ambiental da IA. Desta forma, faz-se necessário um estudo que auxilie a estruturar um modelo de escolha do melhor modelo, que não considere apenas a qualidade das respostas, mas também o tempo de resposta, o custo computacional e o consumo energético.

Partindo dessa premissa, o estudo buscou oferecer um mecanismo estruturado, reprodutível e justificável para apoiar a escolha do modelo mais adequado a cada cenário de uso, explicitando os *trade-offs* entre desempenho, custo e sustentabilidade na aplicação de modelos de linguagem à educação.

Para atingir o objetivo, esta pesquisa adotou uma abordagem experimental comparativa, inserida na lógica da *Design Science Research* (DSR), articulando uma dimensão experimental e uma dimensão de projeto do artefato. Foram avaliados 21 modelos (13 versões base e 8 versões submetidas a ajuste fino), englobando arquiteturas de diferentes portes, abrangendo faixas de tamanho em parâmetros (pequeno, médio e grande) que vão de 1 bilhão até 70 bilhões de parâmetros. A avaliação utilizou 240 questões de múltipla escolha, sendo elas 120 originais elaboradas por professores especialistas e 120 não originais extraídas das edições de 2022, 2023 e 2024 do ENEM, distribuídas em oito disciplinas do ensino médio (português, inglês, geografia, história, matemática, química, física e biologia) e em três níveis de dificuldade (fácil, médio e difícil). As métricas coletadas a partir da execução de cada pergunta em cada modelo alimentaram uma função de utilidade baseada em aproximação de perfis (*Profile Matching*), na qual a seleção do modelo ideal corresponde à minimização da distância euclidiana entre o perfil-alvo do cenário e o desempenho empírico normalizado de cada modelo.

Os resultados obtidos nesta pesquisa indicam que, embora os SLMs de porte médio sejam competitivos em termos de eficiência, os LLMs ainda apresentam desempenho superior em tarefas de maior complexidade, com vantagem clara nas questões de nível Difícil e sem evidências de equivalência estatística nos níveis Fácil e Médio. Ao mesmo tempo, os SLMs mostraram ganhos expressivos de latência e consumo energético, reforçando

seu potencial em cenários alinhados aos princípios da *Green AI*; contudo, dentro da faixa de porte analisada, o aumento de parâmetros não produziu crescimento exponencial de custo, mas sim uma tendência aproximadamente linear.

A análise dos *trade-offs* evidenciou que o ajuste fino, nas condições avaliadas, não trouxe ganho consistente de acurácia, mas elevou significativamente a verbosidade, a latência e o consumo energético, sugerindo um risco de sobreajuste nos modelos. Por fim, a validação do *framework* demonstrou que não existe modelo universalmente superior e que a escolha do melhor modelo deve ser contextual e baseada no equilíbrio entre acurácia, latência, custo computacional e energético, sempre de acordo com os objetivos do cenário proposto.

A principal contribuição deste trabalho é o próprio artefato: um *framework* multicritério, reprodutível e fundamentado empiricamente, capaz de transformar resultados experimentais em recomendações contextualizadas para a seleção de modelos de linguagem em cenários educacionais. Como contribuições adicionais, destacam-se a construção de uma base empírica que articula simultaneamente acurácia, desempenho, custo e sustentabilidade entre modelos, raramente analisados de forma conjunta na literatura. Além disso, essa pesquisa evidencia a escolha por modelos não treinados (*zero-shot*) e que a seleção estratégica de SLMs para tarefas de menor complexidade constitui a via mais pragmática e sustentável para a democratização da inteligência artificial na educação. Por fim, há um alerta empírico quanto aos riscos do ajuste fino indiscriminado, que pode comprometer a eficiência operacional e energética sem retorno proporcional em qualidade.

Apesar das contribuições, o estudo apresenta algumas limitações. As medições foram realizadas em ambiente de inferência local com modelos convertidos para o formato GGUF e quantizados em Q8 nas mesmas condições de hardware, e isso pode condicionar diretamente as medidas de latência, custo computacional e consumo energético, restringindo assim a generalização dos valores absolutos obtidos. Em relação ao treinamento, o ajuste fino foi conduzido em uma GPU NVIDIA T4, utilizando um ambiente de nuvem com restrição de recursos, com a técnica QLoRA/LoRA e uma única época de treinamento sobre uma amostra de 42 mil dados em diferentes idiomas, o que, somado à qualidade dos dados, pode ter influenciado negativamente o desempenho final dos modelos. Em relação ao idioma, todas as perguntas da base de questões estavam em português, sem a inclusão de itens em outros idiomas, o que pode ter influenciado o desempenho dos modelos. Além disso, a avaliação restringiu-se a modelos de acesso aberto executados de forma local, o que não contempla modelos proprietários acessíveis apenas por API. Por fim, a métrica de acurácia adotada limitou-se à corretude objetiva em questões de múltipla escolha, não capturando aspectos qualitativos das respostas, como clareza explicativa e adequação pedagógica.

A partir dessas limitações, apresentam-se diferentes abordagens para trabalhos

futuros. Recomenda-se a ampliação da base de avaliação, em número de questões, disciplinas e dificuldades, bem como a incorporação de métricas qualitativas que avaliem a qualidade pedagógica das explicações, e não apenas a corretude objetiva. Sugere-se, também, a investigação de outras estratégias de ajuste fino, quantização e otimização de inferência em SLMs, a fim de verificar se é possível obter ganhos cognitivos sem o aumento desproporcional de custos observado neste estudo. O *framework* também pode ser estendido com esquemas de ponderação adaptativa entre critérios, com a inclusão de modelos proprietários para fins comparativos e com a ampliação das dimensões de custo, abrangendo, por exemplo, consumo hídrico e custo financeiro de operação. Por fim, mostra-se promissora a validação do artefato em Ambientes Virtuais de Aprendizagem reais, com usuários em uso contínuo, de modo a confrontar as recomendações do *framework* com a experiência prática de estudantes e educadores. Outra frente relevante consiste na extensão do experimento para dispositivos embarcados e de borda (*edge computing*), tais como Raspberry Pi, NVIDIA Jetson, tablets e *smartphones*, nos quais a execução dos SLMs quantizados se torna particularmente pertinente. Essa investigação permitiria verificar empiricamente as recomendações do *framework* sob restrições reais de infraestrutura educacional em contextos de democratização do acesso à IA, além de refinar os patamares das faixas de utilização para dispositivos de menor capacidade computacional.

Em síntese, ao explicitar e gerenciar os *trade-offs* inerentes à adoção de inteligência artificial em larga escala, o *framework* proposto contribui para uma incorporação mais eficiente, responsável e sustentável dessas tecnologias em ambientes educacionais.

7 Uso de Inteligência Artificial

Neste trabalho, foram utilizadas as ferramentas Claude (Anthropic) e ChatGPT (OpenAI), com o objetivo de apoiar atividades específicas relacionadas à elaboração do texto e à produção de elementos visuais. O uso da inteligência artificial esteve restrito às atividades de: apoio na elaboração de um esboço inicial da estrutura do trabalho; auxílio na revisão dos textos produzidos, abrangendo aspectos gramaticais e ortográficos, bem como o aprimoramento da clareza, concisão e organização textual; e geração das Figuras 1, 2, 3, 4 e 5.

Todas as ideias, argumentos, análises, resultados e conclusões apresentadas neste trabalho são de minha autoria e responsabilidade.

Referências

AHMED, A.; IQBAL, M. T.; JAMIL, M. A comparative analysis of power consumption while using open-source and proprietary media players. *European Journal of Electrical Engineering and Computer Science*, v. 8, n. 5, p. 1–5, 2024. Citado na página 64.

ALI, W. et al. Consensus or controversy: Examining ai's impact on academic integrity, student learning, and inclusivity within higher education environments. In: IEEE. *2024 2nd International Conference on Cyber Resilience (ICCR)*. [S.l.], 2024. p. 01–05. Citado 5 vezes nas páginas 16, 20, 21, 31 e 32.

ALIER, M. et al. Lamb: An open-source software framework to create artificial intelligence assistants deployed and integrated into learning management systems. *Computer Standards & Interfaces*, Elsevier, v. 92, p. 103940, 2025. Citado 4 vezes nas páginas 17, 19, 31 e 32.

ANISUZZAMAN, D. et al. Fine-tuning large language models for specialized use cases. *Mayo Clinic Proceedings: Digital Health*, Elsevier, v. 3, n. 1, p. 100184, 2025. Citado na página 28.

ARGERICH, M. F.; PATIÑO-MARTÍNEZ, M. Measuring and improving the energy efficiency of large language models inference. *IEEE Access*, IEEE, v. 12, p. 80194–80207, 2024. Citado na página 27.

BARTELLE, L. B.; MEDEIROS, L. F. de. Os assistentes virtuais nos ambientes virtuais de aprendizagem: uma revisão sistemática de literatura. *EaD em Foco*, v. 14, n. 1, p. e2128–e2128, 2024. Citado na página 20.

BELTON, V.; STEWART, T. *Multiple criteria decision analysis: an integrated approach*. [S.l.]: Springer Science & Business Media, 2012. Citado 4 vezes nas páginas 22, 39, 45 e 48.

BERNARDO, F. et al. Avaliaçao do consumo de energia e o impacto da emissao de co2 para algoritmos de inteligência artificial. In: SBC. *Brazilian e-Science Workshop (BreSci)*. [S.l.], 2020. p. 81–88. Citado na página 36.

CHANDRASEKARAN, B.; NAIR, S. G. S. Unleashing the potential of chatgpt in education: Revolutionizing learning and collaboration. *IEEE Potentials*, v. 43, n. 6, p. 13–22, 2024. Citado na página 18.

CHANGYONG, F. et al. Log-transformation and its implications for data analysis. *Shanghai archives of psychiatry*, v. 26, n. 2, p. 105, 2014. Citado 2 vezes nas páginas 69 e 71.

CHEN, L. et al. Benchmarking large language models on homework assessment in circuit analysis. *International Journal of Artificial Intelligence in Education*, Springer, p. 1–62, 2025. Citado 2 vezes nas páginas 21 e 36.

CHOI, S.; KIM, H. The impact of a large language model-based programming learning environment on students' motivation and programming ability. *Education and Information Technologies*, Springer, p. 1–30, 2024. Citado 4 vezes nas páginas 13, 18, 19 e 32.

CHOI, S.; KIM, H. The impact of a large language model-based programming learning environment on students' motivation and programming ability. *Education and Information Technologies*, Springer, v. 30, n. 6, p. 8109–8138, 2025. Citado na página 35.

CLOUD, A. *Qwen2 Technical Report*. 2024. Accessed on Oct 25, 2025. Disponível em: <https://huggingface.co/Qwen/Qwen2-7B>. Citado na página 57.

CLOUD, A. *Qwen 3 Technical Report*. 2025. Accessed on Oct 25, 2025. Disponível em: <https://github.com/QwenLM/Qwen3>. Citado na página 58.

COBBE, K. et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. Citado 2 vezes nas páginas 42 e 61.

DETTMERS, T. et al. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, v. 36, p. 10088–10115, 2023. Citado na página 59.

ECCARD, W. T. de C.; CAVALCANTE, J. S. C. et al. Inteligência artificial e sustentabilidade: Desafios regulatórios e impacto ambiental. *REVISTA DA FACULDADE DE DIREITO DA UFMG*, v. 86, p. 225–248, 2025. Citado 3 vezes nas páginas 14, 25 e 33.

FAN, A.; LEWIS, M.; DAUPHIN, Y. Hierarchical neural story generation. p. 889–898, 2018. Citado na página 30.

FANG, Q.; OBERSKI, D. L.; NGUYEN, D. Patch! psychometrics-assisted benchmarking of large language models: A case study of proficiency in 8th grade mathematics. *arXiv e-prints*, p. arXiv–2404, 2024. Citado na página 21.

FERNANDEZ, J. et al. Energy considerations of large language model inference and efficiency optimizations. *arXiv preprint arXiv:2504.17674*, 2025. Citado na página 25.

FERNANDEZ, J. et al. Energy considerations of large language model inference and efficiency optimizations. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. [S.l.: s.n.], 2025. p. 32556–32569. Citado na página 27.

GERGANOV, G. *llama.cpp*. 2023. <https://github.com/ggerganov/llama.cpp>. Citado na página 62.

GOOGLE. *Gemma 3 model overview*. 2025. Accessed on Sep 16, 2025. Disponível em: <https://ai.google.dev/gemma/docs/core>. Citado 2 vezes nas páginas 57 e 58.

Google Research. *Google Colaboratory*. 2024. <https://colab.research.google.com/>. Citado na página 59.

HARRIS, T.; YANG, Z.; HARDIN, J. W. Modeling underdispersed count data with generalized poisson regression. *The Stata Journal*, SAGE Publications Sage CA: Los Angeles, CA, v. 12, n. 4, p. 736–747, 2012. Citado 2 vezes nas páginas 69 e 71.

HEVNER, A. R. et al. Design science in information systems research. *Management Information Systems Quarterly*, v. 28, n. 1, p. 6, 2008. Citado 2 vezes nas páginas 23 e 39.

HEWITT, J.; MANNING, C. D.; LIANG, P. Truncation sampling as language model desmoothing. *arXiv preprint arXiv:2210.15191*, 2022. Citado na página 63.

HOLTZMAN, A. et al. The curious case of neural text degeneration. 2020. Citado na página 30.

HORITA, F. E. A.; NETO, V. V. G.; SANTOS, R. P. dos. Design science research em sistemas de informação e engenharia de software: Conceitos, aplicações e trabalhos futuros. *I Jornada Latino-Americana de Atualização em Informática ANAIS*, p. 192, 2018. Citado na página 39.

HU, E. J. et al. Lora: Low-rank adaptation of large language models. *Iclr*, v. 1, n. 2, p. 3, 2022. Citado na página 61.

HUANG, J. et al. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022. Citado na página 18.

Hugging Face. *Transformer Reinforcement Learning (TRL)*. 2024. <https://huggingface.co/docs/trl>. Citado 2 vezes nas páginas 59 e 61.

IA, M. *Introducing Llama 3.1: Our most capable models to date*. 2024. Accessed on Oct 25, 2025. Disponível em: <https://ai.meta.com/blog/meta-llama-3-1/>. Citado 2 vezes nas páginas 57 e 58.

JI, J.-y. J.; KUMAR, R. *Gemma explained: What's new in Gemma 2*. 2024. Accessed on Oct 25, 2025. Disponível em: <https://developers.googleblog.com/gemma-explained-new-in-gemma-2/>. Citado na página 56.

JUNIOR, U. G. et al. Sistemas multiagente e large language model: estudo de caso utilizando as ferramentas lm studio e langgraph. In: SBC. *Workshop-Escola de Sistemas de Agentes, seus Ambientes e Aplicações (WESAAC)*. [S.l.], 2025. p. 250–261. Citado na página 63.

KEENEY, R. L.; RAIFFA, H. *Decisions with multiple objectives: preferences and value trade-offs*. [S.l.]: Cambridge university press, 1993. Citado 3 vezes nas páginas 22, 39 e 45.

LAI, G. et al. RACE: Large-scale ReAding comprehension dataset from examinations. In: PALMER, M.; HWA, R.; RIEDEL, S. (Ed.). *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 785–794. Disponível em: <https://aclanthology.org/D17-1082/>. Citado 2 vezes nas páginas 42 e 61.

LAKENS, D. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, Sage Publications Sage CA: Los Angeles, CA, v. 8, n. 4, p. 355–362, 2017. Citado 2 vezes nas páginas 69 e 71.

LHOEST, Q. et al. Datasets: A community library for natural language processing. In: *Proceedings of the 2021 conference on empirical methods in natural language processing: system demonstrations*. [S.l.: s.n.], 2021. p. 175–184. Citado na página 61.

LI, B. et al. Small language models for application interactions: A case study. *arXiv preprint arXiv:2405.20347*, 2024. Citado na página 18.

LIANG, K.-Y.; ZEGER, S. L. Longitudinal data analysis using generalized linear models. *biometrika*, JSTOR, p. 13–22, 1986. Citado 3 vezes nas páginas 69, 71 e 72.

LIMA, D. A. et al. Uma proposta de sistema de aprendizagem com conteúdo gamificado e com reforço guiado por algoritmos bio-inspirados. *Anais do Computer on the Beach*, v. 8, p. 140–149, 2017. Citado na página 13.

LLAMA. *Llama 3.2: Revolutionizing edge AI and vision with open, customizable models*. 2024. Accessed on Sep 16, 2025. Disponível em: <https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices>. Citado na página 55.

LU, P. et al. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, v. 35, p. 2507–2521, 2022. Citado 2 vezes nas páginas 42 e 61.

MANNING, W. G. The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of health economics*, Elsevier, v. 17, n. 3, p. 283–295, 1998. Citado 2 vezes nas páginas 69 e 71.

MARCO, G.; RELLO, L.; GONZALO, J. Small language models can outperform humans in short creative writing: a study comparing slms with humans and llms. *URL https://arxiv. org/abs/2409.11547*, 2024. Citado 2 vezes nas páginas 24 e 35.

MARTINS, J. P. F. Tomada de decisão multicritério para aumento da sustentabilidade no processo de montagem de bicicletas. 2025. Citado na página 22.

MELISA, R. et al. Critical thinking in the age of ai: A systematic review of ai's effects on higher education. *Educational Process: International Journal*, UNIVERSITEPARK Limited. iTOWER Plaza (No61, 9th floor) Merkez Mh Akar Cd No3 . . . , v. 14, p. e2025031, 2025. Citado na página 13.

MISTRALAI. *Mistral 7B*. 2023. Accessed on Sep 17, 2025. Disponível em: <https://mistral.ai/news/announcing-mistral-7b>. Citado na página 57.

MOREIRA, F. P.; LIMA, D. A. Enhancing early elementary literacy education with flippity tool: challenges, opportunities, and solutions. In: SBC. *Workshop de Informática na Escola (WIE)*. [S.l.], 2023. p. 309–320. Citado na página 13.

NACIMENTO, J. R. d. Exploração de técnicas de engenharia de prompt para aprimorar os resultados do uso de llm no tcmrio. Universidade Federal do Rio Grande do Norte, 2024. Citado na página 29.

NEUMANN, A. T. et al. An llm-driven chatbot in higher education for databases and information systems. *IEEE Transactions on Education*, IEEE, 2024. Citado na página 34.

NGUYEN, T. H. et al. Min-p sampling for creative and coherent language model generation. 2025. Citado 2 vezes nas páginas 30 e 63.

NUTALAPATI, H.; VELMURUGAN, S.; TIGLAO, N. M. Coding buddy: An adaptive ai-powered platform for personalized learning. In: IEEE. *2024 International Symposium on Networks, Computers and Communications (ISNCC)*. [S.l.], 2024. p. 1–6. Citado na página 34.

OH, S. et al. Explore public's perspectives on generative ai in computer science (cs) education: A social media data analysis. In: IEEE. *2024 IEEE Frontiers in Education Conference (FIE)*. [S.l.], 2024. p. 1–9. Citado na página 15.

OPENAI. *GPT-OSS-20B Technical Report*. 2025. Accessed on Oct 25, 2025. Disponível em: <https://huggingface.co/openai/gpt-oss-20b>. Citado na página 57.

OUYANG, F. et al. Comparing the effects of instructor manual feedback and chatgpt intelligent feedback on collaborative programming in china's higher education. *IEEE Transactions on Learning Technologies*, IEEE, 2024. Citado na página 31.

PECHER, B.; SRBA, I.; BIELIKOVA, M. Comparing specialised small and general large language models on text classification: 100 labelled samples to achieve break-even performance. *arXiv preprint arXiv:2402.12819*, 2024. Citado na página 24.

PEFFERS, K. et al. A design science research methodology for information systems research. *Journal of management information systems*, Taylor & Francis, v. 24, n. 3, p. 45–77, 2007. Citado 2 vezes nas páginas 23 e 39.

PEREIRA, V.; BASILIO, M. P.; SANTOS, C. H. T. Enhancing decision analysis with a large language model: pydecision a comprehensive library of mcda methods in python. *Journal of Modelling in Management*, Emerald Publishing Limited, v. 21, n. 2, p. 481–521, 2026. Citado na página 37.

PINHEIRO, J. C.; BATES, D. M. *Mixed-effects models in S and S-PLUS*. [S.l.]: Springer, 2000. Citado 3 vezes nas páginas 69, 71 e 72.

PODDAR, S. et al. Towards sustainable nlp: Insights from benchmarking inference energy in large language models. *arXiv preprint arXiv:2502.05610*, 2025. Citado 2 vezes nas páginas 24 e 25.

REZENDE, G.; DINIZ, A.; DRUMMOND, I. Use of small and large language models for personalization and engagement in virtual learning environments: A systematic literature review. In: . [S.l.: s.n.], 2025. Citado na página 13.

SAATY, T. L. et al. The analytic hierarchy process. McGraw-Hill, New York, 1980. Citado 2 vezes nas páginas 22 e 45.

SACHETE, A. dos S. et al. Você se torna responsável por aquilo que gera: análise semântica e temporal de modelos de linguagem. *RENOTE*, v. 23, n. 1, p. 346–356, 2025. Citado 2 vezes nas páginas 29 e 62.

SAMSI, S. et al. From words to watts: Benchmarking the energy costs of large language model inference. In: IEEE. *2023 IEEE High Performance Extreme Computing Conference (HPEC)*. [S.l.], 2023. p. 1–9. Citado na página 24.

SCHWARTZ, R. et al. Green ai. *Communications of the ACM*, ACM New York, NY, USA, v. 63, n. 12, p. 54–63, 2020. Citado 3 vezes nas páginas 26, 27 e 39.

SENANAYAKE, S.; KARUNANAYAKA, K.; EKANAYAKE, K. Review on ai assistant systems for programming language learning in learning environments. In: IEEE. *2024 8th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI)*. [S.l.], 2024. p. 1–6. Citado na página 16.

STABILITY. *Introducing the Stable LM 2 1.6B*. 2024. Accessed on Sep 16, 2025. Disponível em: <https://stability.ai/news/introducing-stable-lm-2>. Citado na página 56.

STRUBELL, E.; GANESH, A.; MCCALLUM, A. Energy and policy considerations for deep learning in nlp. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*. [S.l.: s.n.], 2019. p. 3645–3650. Citado na página 26.

TRINH, T. H.; DAI, A. M. Taking the temperature of language models. *arXiv preprint arXiv:2208.03188*, 2022. Citado na página 63.

TRUFINESCU, A. *Discover the New Multi-Lingual, High-Quality Phi-3.5 SLMs*. 2024. Accessed on Sep 16, 2025. Disponível em: <https://techcommunity.microsoft.com/blog/azure-ai-foundry-blog/discover-the-new-multi-lingual-high-quality-phi-3-5-slms/4225280>. Citado na página 56.

TSCHAND, A. et al. Mlperf power: Benchmarking the energy efficiency of machine learning systems from μwatts to mwatts for sustainable ai. In: IEEE. *2025 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. [S.l.], 2025. p. 1201–1216. Citado 2 vezes nas páginas 26 e 27.

VEMULA, S. Enriching python programming education with generative ai: Leveraging large language models for personalized support and interactive learning. In: IEEE. *2024 IEEE Frontiers in Education Conference (FIE)*. [S.l.], 2024. p. 1–8. Citado na página 13.

WHITE, J. et al. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023. Citado 2 vezes nas páginas 29 e 62.

WOLF, T. et al. Transformers: State-of-the-art natural language processing. In: *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. [S.l.: s.n.], 2020. p. 38–45. Citado na página 59.

WU, X.-K. et al. Llm fine-tuning: Concepts, opportunities, and challenges. *Big Data and Cognitive Computing*, MDPI, v. 9, n. 4, p. 87, 2025. Citado na página 28.

YADAV, J.; ARORA, S.; KASHYAP, N. Assessing chatgpt's influence on productivity and learning outcomes in higher education: A comprehensive analysis. In: IEEE. *2025 2nd International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*. [S.l.], 2025. p. 428–437. Citado 3 vezes nas páginas 15, 16 e 17.

ZHANG, P. et al. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024. Citado na página 19.