

Transformando evidências da TRI em *feedback* formativo com LLMs: um estudo com microdados do ENEM

Marcelo Lira¹, Edwin Monteiro², Gabriel de Souza Leitão³,
Horácio Fernandes de Oliveira², Raimundo da Silva Barreto²

¹Universidade do Estado do Amazonas (UEA)
Manaus – AM – Brasil

²Instituto de Computação – Universidade Federal do Amazonas (UFAM)
Manaus – AM – Brasil

³Instituto Federal de Educação, Ciência e Tecnologia do Amazonas (IFAM)
Manaus – AM – Brasil

mhdal.snf23@uea.edu.br, gabriel.leitao@ifam.edu.br

{edwin, horacio, rbarreto}@icomp.ufam.edu.br

Abstract. *This paper proposes a method for generating intelligent formative feedback in multiple-choice educational assessments, aiming to transform assessment results into personalized study guidance and support for pedagogical decision-making. As a case study, we use the 2022 ENEM microdata through a pipeline based on Item Response Theory (IRT) and the Three-Parameter Logistic Model (3PL), capable of estimating students' abilities, item parameters, and patterns such as learning gaps, unlikely correct answers, unlikely errors, and priority competencies for review. These evidences are structured and provided to an LLM through pedagogically oriented prompts to generate personalized summaries, explanations of difficulties, improvement guidance, and recommendations for learning resources. The main contribution of this work lies in the integration of automated diagnosis with the intelligent generation of feedback and recommendations, expanding the role of educational assessment in supporting student self-regulation and more targeted teaching interventions.*

Resumo. *Este artigo propõe um método para a geração de feedback formativo inteligente em avaliações educacionais de múltipla escolha, com o objetivo de transformar resultados avaliativos em orientações personalizadas de estudo e em subsídios para a tomada de decisão pedagógica. Como estudo de caso, são utilizados os microdados do ENEM 2022, a partir de um pipeline baseado na Teoria da Resposta ao Item (TRI) e no Modelo Logístico de Três Parâmetros (ML3), capaz de estimar habilidades dos estudantes, parâmetros dos itens e padrões como lacunas de aprendizagem, acertos improváveis, erros improváveis e competências prioritárias para revisão. Essas evidências são estruturadas e fornecidas a um LLM, por meio de prompts orientados pedagogicamente, para gerar sínteses personalizadas, explicações sobre dificuldades, orientações de melhoria e recomendações de recursos de aprendizagem. A principal contribuição do trabalho está na integração entre diagnóstico automatizado e geração*

inteligente de feedback e recomendações, ampliando o papel da avaliação educacional no apoio à autorregulação discente e a intervenções docentes mais direcionadas.

1. Introdução

Avaliações educacionais em larga escala, como o Exame Nacional do Ensino Médio (ENEM), desempenham papel central na mensuração do desempenho dos estudantes e na produção de indicadores sobre a educação brasileira. No entanto, embora esses exames gerem grande volume de dados, as devolutivas disponibilizadas a estudantes e professores ainda são, em geral, pouco exploradas como instrumentos de apoio à aprendizagem. O estudante costuma receber uma nota final ou um número de acertos, mas raramente obtém explicações sobre quais habilidades precisam ser retomadas, quais conteúdos devem ser priorizados e quais padrões de resposta indicam lacunas de aprendizagem.

No ENEM, a Teoria da Resposta ao Item (TRI) é utilizada para estimar a proficiência dos participantes de forma mais robusta do que a simples contagem de acertos, pois considera propriedades dos itens, como dificuldade, discriminação e probabilidade de acerto ao acaso [Lord 1952]; [Andrade et al. 2000]. Em particular, o Modelo Logístico de Três Parâmetros (ML3) permite estimar a probabilidade de um estudante acertar determinado item a partir da relação entre sua habilidade latente e os parâmetros da questão [Birnbbaum 1968]; [Baker e Kim 2017]. Apesar desse potencial, a TRI é frequentemente empregada com finalidade classificatória e somativa, sem que seus resultados sejam convertidos sistematicamente em *feedback* pedagógico compreensível e acionável.

Essa limitação revela uma lacuna no uso educacional desses dados. A proficiência estimada, os parâmetros dos itens e as probabilidades de acerto podem apoiar diagnósticos mais ricos: um acerto em item difícil apesar de baixa probabilidade estimada sugere acerto casual ou domínio instável, enquanto um erro em item de alta probabilidade indica desatenção ou lacuna específica. Organizadas adequadamente, essas evidências subsidiam *feedback* formativo e apoiam o professor na identificação de competências prioritárias.

A literatura sobre *feedback* formativo destaca que devolutivas efetivas devem ir além da indicação de certo ou errado, oferecendo informações que ajudem o estudante a compreender seu desempenho e a tomar decisões sobre seus próximos passos de estudo [van der Kleij et al. 2015]; [Carless 2017]; [Carless e Boud 2018]. Ao mesmo tempo, a geração de *feedback* personalizado em larga escala permanece desafiadora, especialmente em contextos com milhares de participantes e grande quantidade de itens. Sistemas automáticos podem reduzir parte dessa dificuldade, mas muitos deles ainda produzem devolutivas predominantemente quantitativas, genéricas ou pouco conectadas a modelos capazes de explicar a relação entre desempenho observado, habilidade estimada e características dos itens.

Nos últimos anos, os Modelos de Linguagem de Grande Escala (LLMs) passaram a ser explorados em tarefas educacionais como geração de explicações, tutoria e *feedback* automático [Guo et al. 2024]; [Silva e Aranha 2024]. Apesar de produzirem textos fluentes e contextualizados, quando usados isoladamente podem gerar comentários plausíveis, porém pouco ancorados em evidências avaliativas. Permanece, assim, o desafio de combinar sua capacidade gerativa com diagnósticos quantitativos derivados da TRI.

Diante desse cenário, este artigo propõe um método para geração automatizada de *feedback* formativo em avaliações educacionais de múltipla escolha, integrando TRI, ML3 e LLMs. O método organiza um *pipeline* que transforma respostas dos participantes em parâmetros de itens, habilidades estimadas, evidências diagnósticas e relatórios em HTML voltados a estudantes e professores.

A principal contribuição está na articulação entre diagnóstico quantitativo e mediação pedagógica. Enquanto muitos sistemas de recomendação educacional dependem de filtragem colaborativa ou histórico de navegação [Baas et al. 2022], a abordagem proposta utiliza a TRI como motor de evidências para orientar a curadoria de Recursos Educacionais Digitais (RED). O LLM atua como mediador, convertendo métricas do ML3 em explicações, orientações de estudo e recomendações alinhadas às competências avaliadas.

Como estudo de caso, são utilizados os microdados do ENEM 2022, com recorte nos estados do Pará e do Paraná, selecionados por representarem contextos educacionais contrastantes segundo indicadores do Índice de Desenvolvimento da Educação Básica (IDEB) [INEP 2022]. As respostas são organizadas em matrizes dicotômicas e combinadas à Matriz de Referência do ENEM, permitindo relacionar itens, competências, habilidades e evidências de desempenho.

Este artigo está organizado da seguinte forma. A Seção 2 apresenta o referencial teórico, abordando avaliação formativa, Teoria da Resposta ao Item, Modelo Logístico de Três Parâmetros e uso educacional de LLMs. A Seção 3 discute trabalhos correlatos sobre *feedback* automático, TRI e recomendação educacional com inteligência artificial. A Seção 4 descreve a metodologia proposta, incluindo o pré-processamento dos microdados, a calibração do ML3 e a geração automatizada de *feedback* com apoio de LLM. A Seção 5 apresenta os resultados e discussões sobre os relatórios gerados e sua utilidade pedagógica. Por fim, a Seção 6 apresenta as conclusões, limitações e perspectivas de continuidade da pesquisa.

2. Referencial Teórico

Esta seção fundamenta o método proposto: avaliação formativa, Teoria da Resposta ao Item (TRI), Modelo Logístico de Três Parâmetros (ML3) e uso de LLMs como camada de mediação pedagógica.

2.1. Avaliação Formativa e *Feedback* Educacional

A avaliação educacional pode assumir função somativa, quando sintetiza o desempenho ao final de uma etapa, ou função formativa, quando acompanha o percurso de aprendizagem e orienta ações de melhoria durante o processo educativo [Santos 2016]. Na perspectiva formativa, a avaliação deve produzir informações que ajudem o estudante a compreender lacunas de aprendizagem e decidir próximos passos de estudo, ao mesmo tempo em que apoia o professor na retomada de conteúdos e no planejamento de intervenções.

O *feedback* é elemento central nesse processo. De acordo com a literatura, devolutivas podem variar desde formas simples, que apenas indicam acerto ou erro, até formas elaboradas, que incluem explicações, justificativas, recomendações e orientações para estudo [van der Kleij et al. 2015]. Este trabalho se alinha à perspectiva de *feedback* elaborado ao transformar dados de desempenho em orientações interpretáveis, indicando

possíveis razões para erros, sinais de domínio instável, competências prioritárias e recursos educacionais digitais relacionados ao conteúdo avaliado.

A efetividade do *feedback* depende também da capacidade do estudante de compreendê-lo e utilizá-lo. Esse aspecto é discutido por meio do conceito de literacia em *feedback*, que envolve interpretar a devolutiva recebida, fazer julgamentos sobre o próprio desempenho e adotar ações concretas para melhorar a aprendizagem [Carless 2017]; [Carless e Boud 2018]. Assim, sistemas automáticos de *feedback* precisam produzir mensagens claras, específicas e acionáveis, evitando devolutivas genéricas ou excessivamente técnicas.

2.2. Teoria da Resposta ao Item e ML3

A Teoria da Resposta ao Item (TRI) reúne modelos estatísticos que estimam características latentes dos respondentes, como a proficiência, a partir das propriedades dos itens [Lord 1952]; [Andrade et al. 2000]. Diferentemente da Teoria Clássica dos Testes, baseada na contagem de acertos, a TRI relaciona o desempenho do estudante às propriedades dos itens. Por isso, na larga escala do ENEM, dois estudantes com o mesmo número de acertos podem ter proficiências distintas conforme os itens acertados.

Entre os modelos da TRI, este trabalho utiliza o Modelo Logístico de Três Parâmetros (ML3), associado aos modelos de traço latente propostos por [Birnbbaum 1968]. O ML3 estima a probabilidade de um estudante acertar um item considerando três parâmetros: discriminação, dificuldade e acerto ao acaso. A probabilidade de o estudante j acertar o item i , dada sua habilidade latente θ_j , é expressa pela Equação 1.

$$P(U_{ij} = 1|\theta_j) = c_i + (1 - c_i) \frac{1}{1 + e^{-Da_i(\theta_j - b_i)}} \quad (1)$$

Nessa equação, U_{ij} representa a resposta ao item, θ_j a habilidade do estudante, a_i a discriminação, b_i a dificuldade, c_i a probabilidade de acerto ao acaso e D um fator de escala, geralmente fixado em 1 ou 1,7 [Baker e Kim 2017]. Esses parâmetros permitem relacionar desempenho observado, dificuldade do item e probabilidade esperada de acerto.

No contexto deste trabalho, os parâmetros do ML3 são utilizados não apenas para estimar proficiências, mas também para gerar evidências diagnósticas. A relação entre habilidade estimada, dificuldade do item e probabilidade prevista de acerto permite identificar situações em que a resposta observada está alinhada ou desalinhada ao esperado pelo modelo. Essas situações subsidiam a geração de *feedback* formativo e a interpretação pedagógica dos resultados.

2.3. Evidências da TRI para *Feedback*

A aplicação da TRI permite derivar evidências úteis para a construção de *feedback* formativo. Neste trabalho, três padrões são especialmente relevantes: lacunas de aprendizagem, identificadas quando o estudante apresenta desempenho frágil em itens associados a determinadas competências ou habilidades; acertos improváveis, quando o estudante acerta um item apesar de baixa probabilidade estimada de sucesso; e erros improváveis, quando erra um item apesar de alta probabilidade estimada de acerto. Esses padrões não devem ser interpretados de forma determinística, mas funcionam como sinais diagnósticos para orientar revisão de conteúdos, análise de dificuldades e planejamento pedagógico.

2.4. LLMs e Recomendação de Recursos Educacionais Digitais

Modelos de Linguagem de Grande Escala, ou Large Language Models (LLMs), são modelos computacionais capazes de realizar tarefas de compreensão e geração de linguagem natural. Em contextos educacionais, esses modelos têm sido explorados em atividades como tutoria inteligente, geração de explicações, criação de questões e produção de *feedback* personalizado [Guo et al. 2024].

Uma das principais vantagens dos LLMs é transformar informações estruturadas em textos adaptados ao contexto e ao público-alvo. No caso do *feedback* educacional, isso permite converter indicadores técnicos, como habilidade estimada, dificuldade do item e probabilidade de acerto, em mensagens compreensíveis para estudantes e professores. Entretanto, quando utilizados sem dados de apoio ou restrições adequadas, esses modelos podem produzir respostas genéricas ou pouco verificáveis.

Neste trabalho, os LLMs são tratados como camada de mediação pedagógica, não como fonte primária do diagnóstico. O diagnóstico é produzido pela TRI; o LLM recebe essas evidências estruturadas e as transforma em explicações, relatórios em HTML e recomendações de recursos educacionais digitais. Esse uso é coerente com estudos recentes sobre LLMs em recomendação educacional [Ma et al. 2025], mas difere por condicionar a recomendação à competência e à habilidade associadas ao item respondido, priorizando alinhamento curricular e utilidade para estudo.

3. Trabalhos Correlatos

O AutoFeedback, proposto por [Guo et al. 2024], é um sistema multiagente baseado em inteligência artificial generativa que fornece *feedback* automático a respostas textuais de estudantes. No cenário brasileiro, [Silva e Aranha 2024] também exploraram LLMs para *feedback* formativo em programação, demonstrando a eficácia dessa tecnologia na mediação da aprendizagem progressiva. Embora esses trabalhos mostrem o potencial de modelos generativos, o foco reside majoritariamente em respostas abertas ou código, e não na interpretação de padrões de resposta em avaliações de múltipla escolha via TRI.

Ainda no âmbito da múltipla escolha, [Attali et al. 2016] investigaram os efeitos do *feedback* imediato com possibilidade de revisão em questões de múltipla escolha e questões abertas. O estudo reforça a relevância do *feedback* no processo avaliativo, mas também sugere que devolutivas centradas apenas na resposta observada podem ser limitadas. Neste artigo, essa limitação é enfrentada ao relacionar acerto ou erro à habilidade estimada, à dificuldade do item e à probabilidade prevista de acerto.

Já em domínio distinto, [Juniwal 2014] propôs uma abordagem de correção automática e geração de *feedback* no contexto de educação em sistemas ciberfísicos. Embora esteja situado em um domínio diferente do ENEM, o trabalho reforça a importância de mecanismos computacionais para reduzir a carga manual de correção e apoiar professores. Sua contribuição, contudo, não está voltada à análise de itens de múltipla escolha com TRI nem à recomendação de recursos educacionais digitais.

No campo da TRI, [Fujii et al. 2021] utilizaram o Modelo Logístico de Três Parâmetros (ML3) para construir e interpretar escalas de proficiência. Posteriormente, [Monteiro e Barreto 2022] reforçou a aplicabilidade de métodos baseados em TRI para avaliação e *feedback* automático, estendendo trabalhos anteriores de

[Monteiro e Leitão 2021] em educação digital, a aplicabilidade da teoria para além de testes classificatórios. Outros estudos nacionais [Rosa et al. 2024]; [Rauber et al. 2024] aplicaram a TRI para avaliação de coesão textual e escalas de aprendizagem, respectivamente. Ademais, [Romero et al. 2019] analisaram questionários em ambientes virtuais de aprendizagem com modelos da TRI. Embora o modelo de Rasch seja amplamente utilizado por sua simplicidade, o presente trabalho opta pelo ML3 justamente para lidar com o impacto do acerto ao acaso em avaliações de larga escala como o ENEM. Contudo, esses estudos nacionais focam prioritariamente na estimação de proficiência, sem integrar a camada gerativa de LLM para produzir diálogos pedagógicos e recomendações personalizadas.

Além disso, [Monteiro e Barreto 2022] explorou relatórios baseados em TRI com microdados do ENEM e uso do ML3 para identificar lacunas de aprendizagem, acertos improváveis, erros improváveis e itens com maior potencial diagnóstico. O presente artigo parte dessa linha, mas acrescenta uma camada de LLM para produzir explicações mais flexíveis, organizar relatórios em HTML e recomendar recursos educacionais digitais associados às competências e habilidades identificadas.

No eixo da recomendação educacional apoiada por LLMs, [Ma et al. 2025] avaliaram esses modelos na sugestão de cursos em ambientes *MOOC*. No presente método, contudo, a recomendação não parte apenas de preferências ou histórico de navegação: ela é condicionada ao desempenho em um item, à competência e à habilidade da matriz do ENEM e às evidências quantitativas geradas pelo ML3.

Tabela 1. Comparação entre trabalhos correlatos e o método proposto.

Trabalho	Base do diagnóstico	LLM ou recomendação	Diferença para este artigo
Guo et al.	Respostas textuais	LLM para <i>feedback</i>	Não usa TRI nem itens objetivos do ENEM.
Attali et al.	Acerto/erro e revisão	Não usa LLM	Não personaliza a devolutiva por habilidade estimada e parâmetros do item.
Fujii et al.	TRI com ML3	Não usa LLM	Foco em escala de proficiência, não em relatórios formativos.
Romero et al.	TRI com Rasch, ML2 e ML3	Não usa LLM	Foco em análise e ranqueamento de itens.
Ma et al.	Dados de cursos em MOOCs	LLM para recomendação	Recomenda cursos, mas não usa diagnóstico por TRI.
Monteiro e Leitão	TRI com ML3 aplicada ao ENEM	Não usa LLM	Gera relatórios estruturados, sem explicação e recomendação por LLM.
Presente artigo	TRI com ML3 aplicada ao ENEM	LLM para relatório HTML, explicação e recomendação	Integra evidências da TRI, geração de relatórios e recursos digitais.

Em síntese, os trabalhos analisados tratam de forma separada o diagnóstico por TRI, a geração automática de *feedback* e a recomendação com LLM; o método proposto aproxima essas dimensões em um mesmo fluxo, das evidências da TRI ao relatório for-

mativo e à recomendação de recursos alinhados à habilidade a revisar.

4. Metodologia

O método foi organizado como um *pipeline* computacional composto por quatro etapas: (i) pré-processamento dos microdados do ENEM; (ii) calibração do Modelo Logístico de Três Parâmetros (ML3) e estimação das habilidades; (iii) construção de evidências diagnósticas baseadas na relação entre probabilidade de acerto e habilidade; e (iv) geração automatizada de relatórios em HTML com apoio de LLM. As etapas (i) a (iii) herdaram o rigor estatístico do *baseline* de [Monteiro e Barreto 2022], enquanto a etapa (iv) constitui a extensão proposta para mediação pedagógica.

4.1. Pré-processamento dos Microdados

A primeira etapa transformou os microdados brutos do ENEM 2022 em matrizes dicotômicas $N \times I$ (nas quais cada resposta é codificada de forma binária, como acerto (1) ou erro (0)) adequadas à análise TRI. O escopo utiliza o mesmo recorte geográfico e amostral do *baseline*: estudantes do Paraná (PR) e Pará (PA), totalizando 57.703 participantes válidos. Para viabilizar a calibração, selecionou-se um pacote fixo de cadernos de prova (CH1058, LC1067, CN1088 e MT1077), garantindo que todos os examinandos de uma mesma área respondessem aos mesmos itens. Após a limpeza de itens anulados, o conjunto final incluiu 44 itens em CH, 41 em CN, 45 em LC e 43 em MT, após a exclusão de itens sem calibração estável, ou seja, aqueles cuja estimação pelo ML3 não convergiu para valores interpretáveis (discriminação não positiva ou dificuldade divergente). O processamento foi realizado em Python (*pandas*), gerando arquivos CSV que preservam a associação original com a Matriz de Referência do INEP.

4.2. Calibração pelo Modelo Logístico de Três Parâmetros

A calibração dos parâmetros dos itens e a estimação das habilidades seguiram o procedimento técnico documentado em [Monteiro e Barreto 2022]. Utilizou-se o Modelo Logístico de Três Parâmetros (ML3), descrito na Equação 1, para estimar a probabilidade de acerto em função da habilidade do estudante e dos parâmetros de cada item.

A calibração foi realizada na linguagem R, com o pacote *ltm*, que estima os parâmetros dos itens a partir do padrão de acertos e erros de todos os estudantes. O procedimento ajusta o modelo de forma iterativa até encontrar os valores que melhor explicam as respostas observadas.

As habilidades individuais (θ) foram estimadas pelo método *Expected a Posteriori* (EAP) e transformadas para a escala oficial do ENEM (200 a 1000). Os parâmetros a , b , c e as habilidades estimadas alimentam a construção de evidências diagnósticas, incluindo acertos improváveis ($P^* \leq 0,10$), erros improváveis ($P^* \geq 0,90$) e lacunas de aprendizagem, calculadas a partir da probabilidade ajustada $P^* = (P - c)/(1 - c)$.

4.3. Geração Automatizada de *Feedback* com Apoio de LLM

A etapa de geração com LLM teve como objetivo converter as saídas do ML3 em relatórios formativos interpretáveis por estudantes e professores. Diferentemente das etapas anteriores, nas quais a estimação de parâmetros, habilidades e probabilidades é realizada

por procedimentos estatísticos, o LLM foi empregado na mediação textual, na organização visual dos relatórios em HTML e na curadoria assistida de recursos educacionais digitais. Assim, a calibração dos itens, a estimação da habilidade, o cálculo da probabilidade de acerto e a geração das Curvas Características dos Itens (CCI) antecedem a chamada ao modelo de linguagem e constituem entradas estruturadas do processo.

A execução foi organizada como um pipeline automatizado: scripts Python locais processam os dados, geram os artefatos da TRI (CSVs de evidências e gráficos CCI) e acionam a CLI do OpenCode. Esse mecanismo, responsável pela orquestração e não pela estimação, aplica os *prompts* pedagógicos e salva os relatórios em HTML, garantindo que o LLM receba apenas dados já validados pelo motor TRI.

O modelo de linguagem utilizado foi o *Big Pickle*, acessado pela ferramenta *OpenCode*. Ele foi escolhido por processar muitos dados de uma só vez, gerar texto em formato HTML e não ter custo de uso.

O *prompt* especifica uma persona de especialista em psicometria e análise educacional, define regras de preservação da estrutura HTML, delimita quais campos devem ser preenchidos a partir dos valores presentes nos CSVs e orienta a construção do relatório como *feedback* formativo. É fornecido ao modelo um *template* de relatório contendo todas as áreas a serem preenchidas e alguns espaços para livre interpretação do LLM. No relatório do estudante, o modelo interpreta a relação entre habilidade estimada e dificuldade do item, comenta a probabilidade de acerto frente ao parâmetro de acerto ao acaso e sintetiza a competência e a habilidade avaliadas. No relatório do professor, apoia a leitura de gráficos e tabelas agregadas, destacando padrões de desempenho, itens instáveis e ocorrências de respostas inesperadas em função da habilidade estimada.

O detalhamento completo do *prompt*, incluindo a persona, as quatro regras obrigatórias de preenchimento e exemplos de campos, é apresentado na Seção *Disponibilização de Artefatos*.

No tratamento de inconsistências e alucinações, a base de conhecimento dos relatórios é produzida exclusivamente por métodos estatísticos. Os parâmetros dos itens e as habilidades resultam da calibração do ML3; sobre eles são aplicadas transformações adicionais: padronização para a escala do ENEM (200 a 1000), probabilidade ajustada $P^* = (P - c)/(1 - c)$ e classificação de dificuldade dos itens. O LLM é acionado somente na redação textual e na recomendação de recursos, sem recalcular ou inferir grandezas. O risco de alucinação restringe-se, assim, à redação, e não aos dados.

A recomendação de recursos educacionais digitais foi tratada como curadoria assistida por LLM. A partir da competência e da habilidade associadas ao item, o modelo infere tópicos prioritários de estudo e organiza materiais públicos em categorias, como videoaulas, páginas explicativas, PDFs ou slides. Essa recomendação não é apresentada como um sistema de recomendação treinado ou validado empiricamente, mas como recurso exploratório de apoio ao *feedback* formativo, condicionado ao diagnóstico TRI e ao conteúdo curricular indicado pela matriz de referência do ENEM.

Como a seleção de links por LLM pode produzir recursos desalinhados ou indisponíveis, o método impõe restrições de geração. Os materiais devem ser vinculados à competência e à habilidade do item, organizados por tipo de recurso, restritos a fontes públicas e submetidos a inspeção humana antes do uso em larga escala. Os artefatos

públicos gerados pelo *pipeline* estão disponibilizados em página pública¹.

5. Resultados e Discussões

Esta seção apresenta os relatórios formativos gerados a partir dos parâmetros estimados pelo ML3. O objetivo não é apenas visualizar indicadores quantitativos, mas demonstrar como convertê-los em evidências diagnósticas compreensíveis: medidas como habilidade, dificuldade e probabilidade de acerto são interpretadas como indícios de domínio, lacuna, acerto ou erro improvável e instabilidade do item, subsidiando feedback formativo e recomendações alinhadas à matriz do ENEM.

A estrutura dos relatórios em HTML foi definida em um processo iterativo de prototipagem anterior à geração em escala, no qual diferentes versões do *dashboard* foram produzidas e refinadas com o apoio do LLM até se alcançar uma organização adequada à disponibilização e à leitura das evidências. O refinamento norteou-se por princípios consolidados de Interação Humano-Computador: as heurísticas de usabilidade de Nielsen [Nielsen 1994] (com destaque para o reconhecimento em vez da memorização, a visibilidade do estado do sistema e a consistência e padrões), a Teoria da Carga Cognitiva [Sweller 1988], que orienta a redução do esforço de processamento na apresentação da informação, e princípios de visualização e busca de informação em painéis [Shneiderman 1996]. Tais princípios permanecem centrais em revisões recentes sobre concepção e avaliação de painéis de *learning analytics* [Sahin e Ifenthaler 2023]; [Kaliisa et al. 2024], que orientaram a organização das evidências adotada neste trabalho. Uma vez consolidada, a estrutura foi congelada e passou a receber apenas marcadores (*placeholders*) em posições predefinidas, nos quais o LLM injeta as informações diagnósticas. Essa validação restringiu-se à estrutura e à usabilidade do artefato; a avaliação de sua eficácia pedagógica junto a estudantes e professores permanece como trabalho futuro.

5.1. Relatório do Professor

O relatório do professor atua como camada de síntese diagnóstica da turma e dos itens, organizando os parâmetros do ML3 em evidências acionáveis. Acertos improváveis podem sugerir acerto por acaso, fragilidade em distratores ou desalinhamento entre item e habilidade; erros improváveis podem apontar desatenção, ambiguidade na formulação ou necessidade de retomada conceitual.

A Figura 1 ilustra o painel docente, que reúne distribuição de habilidades, diferença entre desempenho observado e esperado e frequência de eventos improváveis. Isso permite priorizar itens para revisão e identificar padrões coletivos: itens com alto parâmetro *c* e muitos acertos improváveis são sinalizados como pontos de atenção.

Como artefato público de demonstração, o relatório do professor foi disponibilizado em uma página estática específica². Essa publicação permite inspecionar o painel interativo gerado, incluindo gráficos por área do conhecimento, indicadores de acertos e

¹https://feedback-formativo-tri.github.io/relatorio_tri-llm/index.html

²https://feedback-formativo-tri.github.io/relatorio_tri-llm/reports/prof_reports/relatorio_professor_PA_v5_10.html#CN

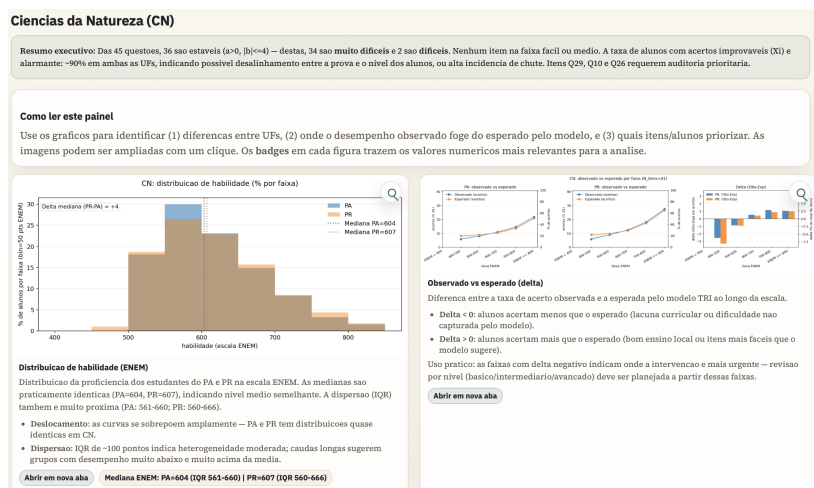


Figura 1. Recorte do relatório docente para Ciências da Natureza, com sínteses de distribuição de habilidade, eventos improváveis, perfis de inconsistência e priorização de itens para revisão pedagógica.

erros improváveis, distribuição de habilidades e alertas de instabilidade dos itens. A disponibilização pública reforça a transparência do processo e facilita a avaliação externa do formato de apresentação dos resultados.

5.2. Relatório do Aluno

No relatório do aluno³, a evidência produzida pela TRI é traduzida em linguagem formativa mediada por LLM. A comparação entre a habilidade estimada (θ) e a dificuldade do item (b) permite que o modelo de linguagem interprete a distância diagnóstica. Quando $\theta < b$, o sistema instrui o LLM a produzir um *feedback* focado em lacunas e recuperação; quando $\theta > b$, o foco reside na consolidação.

A principal contribuição desta extensão é converter eventos e probabilidades do ML3 em devolutiva textual, visual e acionável: enquanto trabalhos anteriores apenas identificavam padrões, o método usa o LLM para explicá-los ao estudante e conectá-los a recomendações de estudo, como na Tabela 2.

A Tabela 2 materializa a extensão proposta através do item 24 de Matemática. Neste caso, o estudante apresentou habilidade 534,0, enquanto a dificuldade do item era 840,86 (b_{ENEM}). A probabilidade de acerto calculada pelo motor TRI ($P(\theta) = 17,14\%$) foi virtualmente idêntica ao parâmetro de acerto ao acaso ($c = 17,13\%$). O LLM interpretou esse dado técnico para produzir um alerta pedagógico sobre a necessidade urgente de revisão dos fundamentos de conjuntos numéricos.

No relatório completo, a CCI é apresentada ao estudante com explicação textual. A Tabela 2 sintetiza o caso: a habilidade fica muito abaixo da dificuldade do item e a probabilidade estimada permanece no piso de acerto ao acaso, preservando a rastreabilidade entre parâmetro, diagnóstico e devolutiva.

A recomendação de materiais constitui a etapa final do fluxo: derivada da com-

³https://feedback-formativo-tri.github.io/relatorio_tri-llm/reports/student_reports/relatorio_210054874307_24_PA_MT.html

Tabela 2. Materialização da Extensão: do diagnóstico estatístico ao feedback para o Item 24 de MT.

Componente	Valor / Descrição
Item / Área	Questão 24 (159) / Matemática e suas Tecnologias
Estatística (Motor)	$\theta = 534,0$; $b = 840,86$; $c = 0,1713$; $P(\theta) = 0,1714$
Diagnóstico (P^*)	Gap de -306,86 pontos ($P^* \approx 0,0$)
Competência (Matriz)	C1: Construir significados para os números naturais, inteiros, racionais e reais.
Output LLM	“O desempenho foi muito baixo, com probabilidade de acerto equivalente ao acaso, indicando necessidade urgente de aprofundar os estudos em conjuntos numéricos.”
Plano de Ação	Tópicos prioritários: Conversão entre representações fracionárias e decimais; Vídeo: “Conjuntos Numéricos - Sandro Curió”.

petência e da habilidade do item, ela não é genérica, mas contextualizada pela matriz de referência do ENEM. A Figura 2 exemplifica essa passagem do diagnóstico para a ação, associando o desempenho no item a tópicos prioritários, recursos para estudo autônomo e pré-visualização do material no próprio relatório.



Figura 2. Feedback do estudante com recomendações de estudo e pré-visualização do recurso selecionado.

Em síntese, a contribuição está na conversão de evidências da TRI em ações interpretáveis: para o professor, organiza sinais para revisão de itens e planejamento; para o estudante, transforma a relação entre habilidade, dificuldade e probabilidade de acerto em feedback e recomendações. Assim, os resultados avaliativos deslocam-se de uma função classificatória para uma função diagnóstica e orientadora da aprendizagem.

6. Conclusão

Os resultados indicam o potencial da integração entre a Teoria da Resposta ao Item (TRI) e Modelos de Linguagem de Grande Escala (LLMs) para reduzir o hiato de interpretação

em avaliações de larga escala. Ao utilizar o ML3 como motor de evidência estatística e o modelo *Big Pickle* (via infraestrutura OpenCode) como camada de mediação pedagógica, foi possível converter parâmetros técnicos de difícil compreensão para o público leigo em relatórios formativos acionáveis, mantendo a rastreabilidade entre a probabilidade de acerto ajustada (P^*) e a explicação textual gerada.

A principal contribuição técnica está na automação de todo o processo: programas em Python preparam os dados e a ferramenta OpenCode aciona o modelo de linguagem, sem intervenção manual. Esse fluxo garante que o LLM não atue como um estimador de proficiência “caixa-preta”, mas sim como um tradutor especializado de evidências já calibradas. Os resultados evidenciam que a identificação de anomalias pedagógicas, como acertos e erros improváveis, ganha utilidade prática ao ser acompanhada de justificativas pedagógicas e recomendações de recursos digitais alinhadas às competências da matriz do ENEM, algo que a TRI isoladamente não oferece.

A TRI sustenta a solução em diferentes volumes de dados, operando de amostras reduzidas a avaliações de larga escala, como o ENEM, que reuniu mais de 4,8 milhões de inscritos em 2025 [INEP 2025]. Além disso, o ML3 generaliza os modelos de um e dois parâmetros: com $c = 0$, reduz-se ao 2PL; com discriminação constante, ao modelo de Rasch [Köhler e Hartig 2021], permitindo adaptar a calibração a amostras menores. A camada de mediação por LLM mantém essa escalabilidade, pois não estima proficiência: opera sobre evidências já calculadas, com custo de inferência previsível e independente do número de relatórios gerados.

A abordagem generaliza-se a qualquer avaliação de múltipla escolha com resposta dicotômica, ou seja, com itens corrigidos de forma binária (acerto ou erro), passível de modelagem pela TRI, incluindo exames de larga escala, simulados e avaliações institucionais. Formatos abertos ou discursivos não são contemplados diretamente; sua incorporação via modelos multidimensionais é uma direção de trabalho futuro.

A principal limitação do trabalho é a ausência de uma avaliação empírica dos relatórios de *feedback* gerados por parte de professores, estudantes e especialistas, que permitiria mensurar diretamente sua qualidade, utilidade e confiabilidade pedagógica. Os resultados aqui apresentados demonstram a viabilidade técnica e o potencial formativo do método, mas não substituem essa validação com o público-alvo. Soma-se a isso a curadoria de recursos externos, que ainda demanda supervisão humana para mitigar riscos de geração não verificada ou obsolescência de links.

Como continuidade, prevê-se um estudo controlado com estudantes e professores, contrastando grupos com e sem acesso aos relatórios, para medir o impacto do *feedback* na aprendizagem e nas decisões docentes, além de um *dashboard* interativo para monitoramento das turmas em tempo real.

Declaração de Uso de IA Generativa

Declara-se o uso de um Modelo de Linguagem de Grande Escala (*Big Pickle*, via ecossistema *OpenCode*) como componente do método proposto, com uma delimitação clara de papéis entre o diagnóstico estatístico e a geração assistida por IA. O diagnóstico foi produzido integralmente por procedimentos psicométricos e estatísticos: a Teoria da Res-

posta ao Item, por meio da calibração do Modelo Logístico de Três Parâmetros, estimou as habilidades dos estudantes, a dificuldade, a discriminação e o acerto ao acaso dos itens; sobre esses parâmetros, foram calculadas estatísticas adicionais e realizada a padronização para a escala do ENEM (200 a 1000), de modo a facilitar a interpretação. A Matriz de Referência do ENEM, com suas competências e habilidades, foi utilizada para identificar os tópicos abordados por cada item; a partir desses tópicos, buscaram-se, por meio do LLM, materiais de estudo compatíveis com o conteúdo exigido. Assim, com base nas evidências previamente calculadas e nos tópicos identificados, o LLM foi empregado para construir os artefatos formativos propriamente ditos, como sínteses textuais, relatórios em HTML e recomendações de recursos (videoaulas, slides e materiais de estudo), destinados a apoiar o estudante e o professor. Ressalta-se que o modelo não estimou parâmetros psicométricos nem substituiu o motor de diagnóstico, e que a redação e a revisão científica do artigo são integralmente humanas.

Disponibilização de Artefatos

O código-fonte do experimento, os scripts de pré-processamento, os dados derivados e os *prompts* completos utilizados estão disponíveis em repositório público⁴, e os relatórios gerados podem ser inspecionados em página pública⁵.

A título de ilustração, o *prompt* do relatório do estudante estabelece a persona de “especialista em psicometria e análise educacional” e define quatro regras obrigatórias: (i) trechos entre parênteses habilitam análise livre no escopo solicitado; (ii) campos entre chaves são substituídos exclusivamente pelos valores presentes nos CSVs; (iii) a ordem dos campos não pode ser alterada; e (iv) a estrutura HTML deve ser rigorosamente preservada. Campos como *{HABILIDADE_ESTIMADA}* e *{PROBABILIDADE_ACERTO}*, bem como o cálculo do *gap* entre a habilidade estimada e a dificuldade do item, são preenchidos a partir das evidências produzidas pela TRI. O *prompt* do relatório do professor segue a mesma lógica e, por sua extensão, permanece disponível no repositório público.

Para mostrar o caráter pedagógico das instruções, segue um trecho do *prompt* do relatório do estudante. O cabeçalho define a persona e as regras de preenchimento:

“Você é um especialista em psicometria e análise educacional. Analise os dados TRI de 3 parâmetros e forneça feedback formativo detalhado que siga o modelo fornecido estritamente. Regras obrigatórias: (1) nos locais com (PARÊNTESES), escreva análises livres no escopo solicitado; (2) nos locais com {CHAVES}, substitua apenas pelo valor correspondente do CSV; (3) não altere a ordem dos campos; (4) preserve rigorosamente a estrutura HTML.”

As instruções entre parênteses concentram o raciocínio pedagógico solicitado ao modelo, sempre ancorado nas evidências da TRI:

- *“Calcule o gap, que é a diferença entre a {HABILIDADE_ESTIMADA_NORMALIZADA} e a {DIFICULDADE_NORMALIZADA}, e infira quanto esforço o aluno deverá empregar para fechá-lo.”*

⁴https://github.com/feedback-formativo-tri/relatorio_tri-llm

⁵https://feedback-formativo-tri.github.io/relatorio_tri-llm/index.html

- “Analise a relação entre $\{PROBABILIDADE_ACERTO\}$ e $\{PROBABILIDADE_ACERTO_ACASO\}$ e indique se é mais provável que o aluno tenha acertado por chute ou por conhecimento.”
- “Usando a competência e a habilidade exigidas pelo item, infira os tópicos prioritários e indique uma estratégia de estudo com intensidade proporcional ao tamanho do gap.”
- “Explique a Curva Característica do Item em nível acessível a um estudante de ensino médio.”

A elaboração dos *prompts* pedagógicos foi conduzida e revisada pelos próprios autores, que reúnem formação e atuação em Inteligência Artificial e em avaliação educacional, assegurando o alinhamento das instruções aos objetivos formativos e à Matriz de Referência do ENEM.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES-PROEX) – Código de Financiamento 001. Este trabalho foi parcialmente financiado pela Fundação de Amparo à Pesquisa do Estado do Amazonas – FAPEAM – por meio do projeto POSGRAD 2026/2027; e também por meio do Edital N. 012/2023, do Programa Startup Áreas Estratégicas.

Referências

- Andrade, D. F. d., Tavares, H. R., e Valle, R. d. C. (2000). *Teoria da Resposta ao Item: conceitos e aplicações*. Associação Brasileira de Estatística, São Paulo.
- Attali, Y., Laitusis, C. C., e Stone, E. (2016). Differences in reaction to immediate feedback and opportunity to revise answers for multiple-choice and open-ended questions. *Educational and Psychological Measurement*, 76(5):787–802.
- Baas, T. et al. (2022). The role of automated feedback in supporting student learning: A systematic review. *Educational Research Review*.
- Baker, F. B. e Kim, S.-H. (2017). *The Basics of Item Response Theory Using R*. Springer, New York, NY.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M. e Novick, M. R., editors, *Statistical Theories of Mental Test Scores*, pages 397–479. Addison-Wesley, Reading, MA.
- Carless, D. (2017). Students' experiences of assessment for learning. In Carless, D., Bridges, S. M., Chan, C. K. Y., e Glofcheski, R., editors, *Scaling up Assessment for Learning in Higher Education*, The Enabling Power of Assessment, pages 113–126. Springer Singapore, Singapore.
- Carless, D. e Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8):1315–1325.
- Fujii, T. R. C., Souza, A. D. P. d., Fürkotter, M., Borgatto, A. F., e Cúri, M. (2021). Estudo sobre construção de escalas com base na teoria da resposta ao item: Avaliação

- de proficiência em conteúdos matemáticos básicos. *Bolema: Boletim de Educação Matemática*, 35(71):1876–1898.
- Guo, S., Latif, E., Zhou, Y., Huang, X., e Zhai, X. (2024). Using generative ai and multi-agents to provide automatic feedback.
- INEP (2022). Índice de desenvolvimento da educação básica (ideb) 2021: Resultados e informações. Technical report, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.
- INEP (2025). Enem 2025 tem mais de 4,8 milhões de inscritos confirmados. Ministério da Educação. Disponível em: <https://www.gov.br/mec>.
- Juniwal, G. (2014). Cpsgrader: Auto-grading and feedback generation for cyber-physical systems education. Master's thesis, University of California, Berkeley. Technical Report No. UCB/EECS-2014-237.
- Kaliisa, R., Misiejuk, K., López-Pernas, S., Khalil, M., e Saqr, M. (2024). Learning analytics dashboards are increasingly becoming about learning and not just analytics — a systematic review. *Education and Information Technologies*, 29:14279–14308.
- Köhler, C. e Hartig, J. (2021). Practical significance of item misfit and its manifestations in constructs assessed in large-scale studies. *Frontiers in Psychology*, 12.
- Lord, F. M. (1952). *A Theory of Test Scores*. Number 7 in Psychometric Monograph. Psychometric Corporation, Richmond, VA.
- Ma, L. et al. (2025). Llm-based course recommendation in moocs.
- Monteiro, E. J. L. B. e Barreto, R. S. (2022). Um método baseado na teoria da resposta ao item para avaliação e feedback automático no contexto do enem. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação (SBIE 2022)*, pages 225–236, Porto Alegre, RS, Brasil. SBC.
- Monteiro, E. J. L. B. e Leitão, G. S. (2021). Um método baseado na teoria da resposta ao item para avaliação e feedback automático no contexto de educação digital. *Revista Brasileira de Informática na Educação (RBIE)*, 29:1006–1031.
- Nielsen, J. (1994). *Usability Engineering*. Morgan Kaufmann, San Francisco.
- Rauber, M. F., von Wangenheim, C. G., et al. (2024). Proposta de escala para avaliação da aprendizagem de machine learning em nível create na educação básica. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação (SBIE 2024)*, Porto Alegre, RS, Brasil. SBC.
- Romero, O. O., Machado, G. M., e Wives, L. K. (2019). Analysis of questionnaires for virtual learning environments based on item response theory. In *Anais do XXX Simpósio Brasileiro de Informática na Educação (SBIE 2019)*, pages 509–518, Porto Alegre, RS, Brasil. Sociedade Brasileira de Computação.
- Rosa, B. A., Oliveira, H., e Mello, R. F. (2024). Combinação de modelos de aprendizado de máquina utilizando teoria de resposta ao item para avaliação de coesão textual em redações no contexto do enem. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação (SBIE 2024)*, Porto Alegre, RS, Brasil. SBC.

- Sahin, M. e Ifenthaler, D. (2023). A checklist to guide the planning, designing, implementation, and evaluation of learning analytics dashboards. *International Journal of Educational Technology in Higher Education*, 20(28).
- Santos, L. (2016). A articulação entre a avaliação somativa e a formativa, na prática pedagógica: uma impossibilidade ou um desafio? *Ensaio: Avaliação e Políticas Públicas em Educação*, 24(92):637–669.
- Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, pages 336–343. IEEE Computer Society.
- Silva, F. G. e Aranha, E. H. S. (2024). Feedback formativo automatizado com llms: Desenvolvimento e análise de um sistema para aprendizagem progressiva em programação. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação (SBIE 2024)*, Porto Alegre, RS, Brasil. SBC.
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.
- van der Kleij, F. M., Feskens, R. C. W., e Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research*, 85(4):475–511.