

**AVALIAÇÃO CRÍTICA DE
RESPOSTAS DE
INTELIGÊNCIA ARTIFICIAL
GENERATIVA POR
ESTUDANTES DO ENSINO
MÉDIO EM QUÍMICA,
FÍSICA, BIOLOGIA E
MATEMÁTICA**

**CRITICAL EVALUATION OF GENERATIVE ARTIFICIAL INTELLIGENCE
RESPONSES BY HIGH SCHOOL STUDENTS IN CHEMISTRY, PHYSICS,
BIOLOGY, AND MATHEMATICS**

Ciências Exatas e da Terra, Ciências Humanas, Ciências Biológicas •

22/09/2026

REGISTRO DOI: [10.70773/revistatopicos/790052529](https://doi.org/10.70773/revistatopicos/790052529)

David Willames do Rego Coimbra¹

Raul Ferreira de Miranda Mendes²

Ana Maria Silva Costa³

Wedson Medeiros Silva Souto⁴

Jorge Soares Menor Filho⁵

Rosiane da Silva Sales⁶

Natanael da Silva Rocha⁷

Alissony Moraes Paiva⁸

Francinildo Costa da Silva Lemos⁹

Weverson Fernandes Batista da Silva¹⁰

Murilo Benicio Araujo Sousa¹¹

Isaac Cesar Pereira Cardoso Junior¹²

Renildo da Silva do Nascimento¹³

Nara Suellen Leoncio Chaves¹⁴

Cleiton James Gomes¹⁵

Rejanne Alves de Melo¹⁶

RESUMO

A inteligência artificial generativa (IAG) já faz parte do cotidiano escolar, mas seu uso não significa que os estudantes consigam avaliar a validade das respostas produzidas por esses sistemas. Esta pesquisa empírica, interdisciplinar e descritiva envolveu 83 estudantes de três turmas do Ensino Médio, uma de cada série, em uma escola pública da zona norte de Teresina, Piauí. O instrumento apresentou 24 situações-problema de Química, Física, Biologia e Matemática, com respostas incorretas, parcialmente plausíveis e itens corretos de controle. O protocolo V5, composto pelas ações de detectar, localizar, verificar, corrigir e justificar, orientou a atividade em articulação com o quadro de Avaliação Crítica de Respostas de Inteligência Artificial (IAC-IA). Nesta análise, foram examinadas a classificação da validade das respostas e a percepção declarada pelos participantes; as respostas abertas do V5/IAC-IA não estavam disponíveis em formato individual para codificação sistemática. A classificação esperada variou de 18,1% a 80,7%. Os maiores percentuais ocorreram nos itens sobre velocidade média (80,7%), fotossíntese (78,3%) e nas respostas corretas de controle. As menores taxas envolveram pH e neutralidade (18,1%), determinação de função por dois pontos (18,1%) e propriedades relacionadas ao ponto de ebulição (21,7%). Na escala de percepção, 81,9% concordaram que uma resposta bem escrita não é necessariamente correta, e 80,7% afirmaram que pretendem verificar as respostas de IAG com mais cuidado. Os resultados descrevem diferenças entre os itens e uma distância entre a cautela declarada e o desempenho observado, sem estabelecer efeito causal ou desenvolvimento individual.

Palavras-chave: inteligência artificial generativa; Ensino Médio; educação STEM; alfabetização em IA; pensamento crítico; argumentação científica; raciocínio matemático.

ABSTRACT

Generative artificial intelligence (GenAI) is already part of everyday school life, but using these tools does not necessarily mean that students can judge the validity of their outputs. This empirical, interdisciplinary, descriptive study involved 83 students from three secondary-school classes, one from each grade level, at a public school in northern Teresina, Piauí, Brazil. The instrument presented 24 problem situations in Chemistry, Physics, Biology, and Mathematics, including incorrect answers, partially plausible responses, and correct control items. The V5 protocol, comprising detect, locate, verify, correct, and justify, guided the activity together with the Critical Evaluation of Artificial Intelligence Responses framework (IAC-IA, Portuguese acronym). The present analysis examined validity classifications and participants' reported perceptions; individual open-ended V5/IAC-IA responses were not available for systematic coding. Expected classification rates ranged from 18.1% to 80.7%. The highest rates occurred for average speed (80.7%), photosynthesis (78.3%), and the correct control items. The lowest rates involved pH and neutrality (18.1%), function determination from two points (18.1%), and boiling-point reasoning (21.7%). On the perception scale, 81.9% agreed that a well-written answer is not necessarily correct, and 80.7% intended to verify GenAI outputs more carefully. The findings describe differences among items and a gap between reported caution and observed performance, without establishing causal effects or individual development.

Keywords: generative artificial intelligence; secondary education; STEM education; AI literacy; critical thinking; scientific argumentation; mathematical reasoning.

1. INTRODUÇÃO

A inteligência artificial generativa alterou rapidamente o ecossistema de produção e circulação de informações. Em ambientes escolares, sistemas capazes de produzir explicações, sínteses, cálculos, argumentos e soluções passaram a oferecer respostas em linguagem natural com baixo custo de interação. A consequência pedagógica não é apenas maior acesso à informação: muda também o tipo de competência necessário para que o estudante diferencie uma resposta plausível de uma resposta válida.

Essa distinção é particularmente importante nas áreas de Ciência, Tecnologia, Engenharia e Matemática (STEM, do inglês Science, Technology, Engineering and Mathematics). Em Química, uma explicação pode empregar terminologia adequada e ainda confundir equilíbrio dinâmico com paralisação da reação. Em Física, uma solução pode citar corretamente uma lei e manipular de modo incorreto as grandezas. Em Biologia, uma resposta pode converter correlação em causalidade ou extrapolar evidências limitadas. Em Matemática, uma conclusão correta pode decorrer de um argumento inválido. Em todos esses casos, a fluência textual funciona como um sinal superficial que não substitui a verificação disciplinar.

Revisões publicadas em 2026 mostram que a evidência sobre IAG em STEM está crescendo, mas permanece heterogênea. Zhao et al. (2026) sintetizaram 61 estudos empíricos e observaram predominância do ensino superior. Boolzen et al. (2026), em revisão e meta-análise de resultados cognitivos em STEM, localizaram 85 estudos elegíveis, 49 deles incluídos na meta-análise, e encontraram heterogeneidade substancial entre os efeitos. Esses resultados recomendam cautela diante de afirmações genéricas de que a IAG,

por si só, melhora a aprendizagem: o desenho da atividade e a qualidade da interação importam.

No campo da alfabetização em IA, Niri et al. (2026) revisaram 39 estudos de educação STEAM e destacaram que as evidências sobre como diferentes dimensões da AI literacy são desenvolvidas em contextos interdisciplinares ainda se encontram fragmentadas. Em Química, Young, Dawood e Lewis (2024) já demonstraram a pertinência de investigar reflexões críticas de estudantes sobre respostas de chatbots. Assim, a lacuna defensável não é a simples existência de avaliação de respostas de IA, mas a investigação sistemática, em estudantes do Ensino Médio, de como estratégias de avaliação crítica operam e eventualmente se transferem entre diferentes domínios STEM.

Este estudo propõe uma inversão pedagógica: a resposta da IAG deixa de ser produto final e passa a ser estímulo para investigação. O estudante é solicitado a detectar, localizar, verificar, corrigir e justificar. A ênfase desloca-se da eficiência de obter uma resposta para a qualidade do raciocínio utilizado para julgá-la. Essa arquitetura permite aproximar alfabetização em IA, alfabetização científica, raciocínio matemático, argumentação e metacognição.

A pergunta central é, portanto, menos tecnológica e mais epistemológica: quando uma máquina apresenta uma resposta convincente, que recursos o estudante mobiliza para decidir se deve aceitá-la? E, quando muda a disciplina, o estudante mantém a mesma estratégia crítica ou depende de formas de conhecimento específicas do domínio?

Figura 1. Modelo conceitual: da resposta pronta à investigação crítica.



A IAG funciona como objeto de análise; o ganho pedagógico depende da atividade cognitiva do estudante.

Fonte: elaboração própria.

2. ESTADO DA ARTE E LACUNA DE PESQUISA

2.1. IAG e Aprendizagem em STEM

A literatura recente converge em um ponto: os efeitos educacionais da IAG são condicionais. A revisão sistemática de Zhao et al. (2026) descreve oportunidades de personalização, inovação pedagógica e ampliação do acesso a recursos, ao mesmo tempo em que identifica riscos de dependência, confiabilidade e questões relacionadas ao uso. A meta-análise de Boolzen et al. (2026) reforça a necessidade de interpretar médias agregadas com cuidado, pois os estudos diferem fortemente quanto a população, tarefa, ferramenta e desenho instrucional. Para o presente estudo, essa heterogeneidade é teoricamente relevante: a variável decisiva pode não ser “usar ou não usar IA”, mas o nível de atividade cognitiva exigido do estudante durante a interação.

2.2. Pensamento Crítico e Mediação Estruturada

Revisões recentes sobre IAG e pensamento crítico sugerem que benefícios cognitivos aparecem com maior consistência quando a tecnologia é inserida em tarefas estruturadas, investigativas e acompanhadas de mediação pedagógica, enquanto usos pouco estruturados podem favorecer terceirização cognitiva. No desenho adotado neste estudo, a IAG não executa a tarefa pelo estudante; sua resposta constitui um artefato a ser examinado, confrontado e justificado à luz de conhecimentos disciplinares.

2.3. Alfabetização em IA e Interdisciplinaridade

A alfabetização em IA envolve dimensões técnicas, críticas e sociais. Neste estudo, o foco está na capacidade de reconhecer que uma resposta produzida por IAG precisa ser analisada e verificada, e não aceita automaticamente. O caráter interdisciplinar permite observar se essa postura se mantém quando mudam as representações, os critérios de evidência e os procedimentos de validação.

A literatura mais recente reforça a necessidade de separar uso de IAG de alfabetização crítica em IAG. Revisões em educação básica apontam expansão rápida das aplicações, mas ainda com forte diversidade de objetivos, desenhos e medidas; estudos sobre alfabetização em IA generativa destacam a avaliação e a incorporação crítica de saídas generativas como competência distinta do simples uso instrumental. Nesse cenário, o presente estudo concentra-se na qualidade do julgamento produzido pelo estudante e nas evidências mobilizadas para sustentá-lo, em vez de assumir que maior exposição à ferramenta implica maior aprendizagem.

Figura 2. Matriz interdisciplinar das formas de raciocínio focalizadas.

Matriz interdisciplinar de raciocínio



Questão analítica: a estratégia crítica se transfere entre domínios ou depende do conhecimento disciplinar?

Fonte: elaboração própria.

2.4. Formulação da Lacuna

A lacuna investigada é delimitada em quatro níveis. Primeiro, há menor concentração de evidências em educação básica do que em ensino superior nas revisões recentes. Segundo, estudos sobre reflexão crítica de respostas de chatbot existem, mas frequentemente se concentram em uma disciplina ou contexto específico. Terceiro, permanece pouco explorada a comparação, no mesmo desenho, entre erros conceituais, quantitativos, lógicos e argumentativos. Quarto, a transferência de uma estratégia de verificação entre disciplinas pode ser tratada como resultado de aprendizagem próprio, e não apenas como efeito colateral.

2.5. Inserção no Contexto Curricular Brasileiro

No Ensino Médio brasileiro, a escolha de Química, Física, Biologia e Matemática não constitui apenas uma conveniência organizacional. A Base Nacional Comum Curricular estrutura Ciências da Natureza e suas Tecnologias a partir de Química, Física e Biologia e mantém Matemática e suas Tecnologias como área própria, ao mesmo tempo em que atribui centralidade a pensamento científico, crítico e

criativo, cultura digital e argumentação. Essa arquitetura curricular oferece um terreno particularmente adequado para investigar se procedimentos de avaliação crítica aprendidos em uma área reaparecem em outra.

Na Matemática, a BNCC enfatiza interpretação de situações, construção de modelos, análise da plausibilidade dos resultados, argumentação consistente e investigação de conjecturas. Em Ciências da Natureza, a ênfase recai sobre interpretação de modelos, investigação, análise de situações-problema e reconhecimento de potencialidades e limitações do conhecimento científico. Assim, pedir ao estudante que analise uma resposta de IAG não é uma atividade periférica ao currículo: quando bem desenhada, pode mobilizar práticas já valorizadas na formação científica e matemática, como justificar, testar, comparar, modelar e revisar conclusões.

A contextualização em Teresina, Piauí foi tratada como contexto empírico, e não como argumento de generalização. O estudo não pretende afirmar que os resultados representam todos os estudantes brasileiros; descreve as condições locais de produção dos dados e discute a possível transferibilidade dos procedimentos pedagógicos para outros contextos, sempre distinguindo o contexto da amostra do alcance das inferências.

3. REFERENCIAL TEÓRICO

3.1. Alfabetização Científica

A alfabetização científica é tratada aqui como capacidade de mobilizar conceitos, modelos e evidências para interpretar afirmações e justificar conclusões. A avaliação de uma saída de IAG

torna-se uma prática de alfabetização científica quando exige comparar a afirmação com conhecimento disciplinar, reconhecer limites de evidência e reconstruir uma explicação.

3.2. Raciocínio Matemático

O raciocínio matemático envolve relações, representações, procedimentos, generalizações e justificativas. A avaliação crítica de uma solução gerada por IA permite distinguir erro de cálculo, erro de estratégia, inferência inválida e conclusão não sustentada. Essa distinção é central porque respostas matemáticas podem chegar ao valor correto por procedimento inadequado.

3.3. Argumentação

A argumentação permite que o estudante explique como chegou ao seu julgamento. A estrutura considerada no estudo inclui uma conclusão sobre a validade da resposta, uma justificativa e um fundamento relacionado ao conteúdo. Nos níveis mais altos da escala de avaliação, o estudante também considera condições, limites e possíveis alternativas.

3.4. Metacognição e Calibração de Confiança

A competência crítica inclui monitorar a própria certeza. Um estudante pode identificar um erro por intuição sem conseguir justificá-lo; outro pode aceitar uma resposta errada porque sua forma textual transmite confiança. No instrumento aplicado, o grau de confiança integrou a estrutura de resposta, permitindo considerar a segurança declarada como dimensão metacognitiva complementar ao julgamento de validade.

3.5. Aparência de Correção e Confiabilidade das Respostas

Uma resposta pode ser fluente, organizada e utilizar vocabulário técnico e, ainda assim, apresentar erro conceitual, cálculo inadequado ou conclusão sem sustentação. Neste estudo, o interesse está em observar em que condições o estudante considera uma informação confiável. A proposta não é incentivar desconfiança automática da tecnologia, mas desenvolver o hábito de verificar as respostas antes de aceitá-las.

Essa distinção amplia a unidade de análise. O julgamento do estudante não se restringe a acertar ou errar a classificação de uma saída de IAG; interessa também a capacidade de localizar o problema, indicar formas de verificação, corrigir a resposta e justificar a decisão. Conceitualmente, uma classificação correta obtida sem fundamento não equivale a uma avaliação criticamente sustentada, razão pela qual o instrumento foi estruturado para tornar visíveis diferentes operações do julgamento.

A noção de autoridade algorítmica é utilizada de forma descritiva para discutir situações em que a fluência, o detalhamento, o uso de fórmulas ou o vocabulário técnico de uma resposta podem funcionar como pistas de credibilidade. O instrumento combinou respostas corretas e incorretas, evitando que a tarefa pudesse ser resolvida pela expectativa de que toda saída atribuída à IAG estivesse errada.

3.6. Transferência Entre Domínios

Transferência é entendida como reutilização de uma estratégia de avaliação em contexto disciplinar diferente. O estudo não pressupõe que essa transferência ocorra automaticamente. Os padrões

observados entre as quatro áreas foram interpretados preservando a possibilidade de que o conhecimento específico de cada domínio limite a generalização da estratégia.

4. QUESTÕES DE PESQUISA E OBJETIVOS

4.1. Questões de Pesquisa

Código	Questão
RQ1	Como se distribui o desempenho dos estudantes na classificação da validade das 24 respostas atribuídas à IAG?
RQ2	Como as taxas de classificação esperada variam entre os itens de Química, Física, Biologia e Matemática?
RQ3	Quais situações apresentam maior e menor frequência de classificação esperada?
RQ4	Os estudantes reconhecem respostas corretas de controle ou adotam uma estratégia de rejeição indiscriminada da IAG?
RQ5	Como os estudantes percebem sua capacidade de verificar, localizar e corrigir respostas de IAG após a atividade?
RQ6	Que relação descritiva emerge entre a consciência declarada sobre a necessidade de cautela e o desempenho observado nos itens?

4.2. Objetivo Geral

Analisar como estudantes do Ensino Médio avaliam criticamente respostas produzidas por IAG em Química, Física, Biologia e Matemática, considerando detecção, localização, verificação, correção, fundamentação, argumentação e confiança declarada.

4.3. Objetivos Específicos

- Caracterizar a amostra e a frequência declarada de uso de ferramentas de IAG.
- Comparar o desempenho na classificação da validade entre disciplinas e tipos de erro.
- Examinar diferenças entre erros diretos, manipulações sutis, concepções alternativas e problemas de raciocínio lógico ou estatístico.
- Descrever o desempenho dos estudantes na classificação da validade das respostas apresentadas.
- Analisar a percepção dos estudantes sobre verificação, correção, justificativa e uso crítico de IAG.
- Integrar, em nível descritivo, o desempenho observado e a percepção declarada após a atividade.
- Discutir implicações pedagógicas para alfabetização crítica em IA no Ensino Médio.
- Documentar um instrumento interdisciplinar replicável para investigação de avaliação crítica de respostas de IAG.

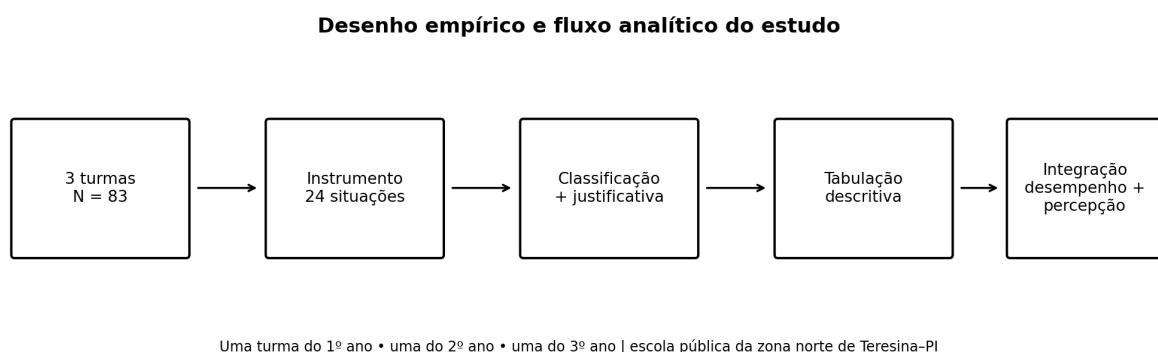
5. MÉTODO

5.1. Desenho do Estudo

Foi realizado um estudo empírico, interdisciplinar, de abordagem quantitativa e caráter descritivo. A investigação buscou caracterizar como estudantes das três séries do Ensino Médio classificaram

respostas atribuídas à IAG em quatro áreas STEM e como perceberam sua própria capacidade de verificação. Como a base disponível está tabulada de forma agregada, os resultados são apresentados por frequências, proporções e médias descritivas, sem inferência causal, comparação individual pré-pós ou codificação sistemática das justificativas abertas.

Figura 3. Desenho empírico e fluxo analítico do estudo.



Fonte: elaboração própria.

5.2. Contexto e Participantes

A pesquisa foi realizada em uma escola pública de Ensino Médio situada na zona norte de Teresina, Piauí. Participaram 83 estudantes de três turmas, uma de cada série: 28 do 1º ano, Turma A (33,7%); 28 do 2º ano, Turma B (33,7%); e 27 do 3º ano, Turma C (32,5%). Essa distribuição permitiu observar diferentes momentos da trajetória escolar. O nome da instituição foi omitido para preservar a confidencialidade dos participantes e do contexto pesquisado.

5.2.1. Critérios de Participação e Caracterização da Amostra

Foram incluídos os estudantes das três turmas que participaram da aplicação e apresentaram respostas válidas, totalizando N = 83. Entre eles, 26 participantes (31,3%) tinham de 14 a 15 anos, 47 (56,6%) tinham de 16 a 17 anos e 10 (12,1%) tinham 18 anos ou mais. Nenhum participante foi excluído em razão do desempenho acadêmico. A análise considerou o conjunto da amostra, sem separar os resultados por turma.

5.2.2. Ambiente de Aplicação e Controle de Contaminação

A aplicação foi realizada em ambiente escolar, com orientação padronizada e estímulos de IAG previamente definidos. O instrumento foi estruturado para que os participantes julgassem respostas corretas, parcialmente corretas ou incorretas e explicitassem localização do problema, estratégia de verificação, correção, justificativa e grau de confiança. Respostas corretas de controle foram incluídas para reduzir a estratégia de rejeitar automaticamente toda saída atribuída à IAG.

5.2.3. Planejamento Amostral e Precisão das Estimativas

A amostra reuniu os 83 participantes efetivamente analisados, distribuídos em três turmas, uma de cada série. O estudo descreve padrões observados nesse contexto escolar específico, sem pretensão de representar todos os estudantes de Teresina ou do Piauí. Como os dados foram organizados de forma agregada, a análise concentrou-se nas frequências, médias e diferenças observadas no conjunto dos participantes.

5.3. Procedimento Pedagógico e Aplicação do Instrumento

Etapa do instrumento	Foco	Ação solicitada ao estudante	Evidência registrada
1	Caracterização	Informar série, faixa etária e frequência de uso de IAG	Perfil agregado da amostra
2	Classificação	Julgar a validade de cada resposta apresentada	Categoria selecionada no Item A
3	Localização	Indicar onde se encontra eventual problema	Registro textual
4	Verificação	Explicar como testaria a resposta	Procedimento ou evidência proposta
5	Correção/justificativa	Reconstruir e fundamentar quando necessário	Resposta argumentada
6	Percepção	Avaliar afirmações sobre uso crítico de IAG	Escala Likert de 1 a 5

A atividade pedagógica foi organizada para que os estudantes diferenciassem uma resposta bem apresentada de uma resposta realmente válida. As tarefas envolveram conferência de conceitos, cálculos e unidades, análise de evidências, justificativas e correção das respostas. O instrumento aplicado ao final reuniu 24 situações-problema distribuídas entre Química, Física, Biologia e Matemática, incluindo quatro respostas corretas de controle (Q6, Q12, Q18 e Q24).

O professor atuou como mediador da verificação, solicitando que os estudantes explicitassem os critérios utilizados para aceitar, rejeitar ou corrigir uma resposta. A mediação privilegiou perguntas sobre

evidências, cálculos, contraexemplos, condições de validade e justificativas, evitando reduzir a atividade a uma simples busca por erros.

5.4. Protocolo V5

O núcleo da atividade foi o protocolo V5: detectar, localizar, verificar, corrigir e justificar. A sequência evitou reduzir a tarefa à simples marcação de “certo” ou “errado” e solicitou a externalização do raciocínio. O instrumento também incluiu uma pergunta metacognitiva sobre o grau de confiança do estudante em sua própria avaliação, registrada em escala previamente definida.

Figura 4. Protocolo V5 solicitado ao estudante diante de uma resposta de IAG.

Protocolo V5 de avaliação crítica



A resposta final só é considerada robusta quando o julgamento é acompanhado por fundamento disciplinar e reconstrução.

Fonte: elaboração própria.

5.5. Padronização do Protocolo de Resposta

A aplicação seguiu orientação comum às três turmas, com manutenção dos mesmos estímulos e da mesma sequência de operações solicitadas aos estudantes. Essa padronização teve como

finalidade reduzir variações procedimentais entre as turmas. Como a base disponível para esta análise é agregada, o estudo não estima efeito causal da mediação nem compara fidelidade de implementação entre aplicadores.

5.6. Instrumento e Estímulos de IAG

5.6.1. Composição do Banco de Situações-Problema

O instrumento contém 24 situações-problema, seis por disciplina, assegurando igual representação de Química, Física, Biologia e Matemática. Os itens foram organizados para contemplar erros conceituais, procedimentais e de raciocínio, além de quatro respostas corretas de controle. Essa composição permitiu observar se os estudantes distinguiam respostas válidas de respostas apenas aparentemente convincentes.

Figura 5. Balanceamento do banco de 24 itens utilizado no estudo.

Balanceamento planejado do banco de 24 itens

	Conceitual	Quantitativo / procedimental	Crítica da IAG
Química	2 itens	2 itens	2 itens
Física	2 itens	2 itens	2 itens
Biologia	2 itens	2 itens	2 itens
Matemática	2 itens	2 itens	2 itens

6 itens por disciplina • 8 itens por modalidade • desenho balanceado antes da validação

Área	Conceitual	Quantitativo/procedimental	Crítica da IAG	Total
Química	2	2	2	6
Física	2	2	2	6
Biologia	2	2	2	6
Matemática	2	2	2	6
TOTAL	8	8	8	24

5.6.2. Geração e Congelamento dos Estímulos de IAG

Para aumentar a reprodutibilidade, os estímulos de IAG foram definidos antes da aplicação e mantidos inalterados durante a coleta. Foram preservados o enunciado apresentado, o contexto fornecido e a classificação prevista quanto à validade. Todos os participantes receberam os mesmos estímulos correspondentes ao instrumento, evitando que alterações posteriores em sistemas generativos interferissem retrospectivamente nas situações analisadas.

5.6.3. Controle de Expectativa de Erro

Os estímulos foram balanceados para impedir a regra informal de que “a IA sempre está errada”. O banco reuniu respostas corretas e respostas com erros ou limitações, incluindo situações em que a plausibilidade textual poderia competir com a análise disciplinar. Os quatro itens corretos de controle permitiram verificar se os participantes eram capazes de reconhecer saídas adequadas, em vez de rejeitar indiscriminadamente respostas atribuídas à IAG.

5.6.4. Tipologia de Respostas

Tipo	Descrição	Risco cognitivo que se pretende testar
A	Correta e bem justificada	Falso positivo: rejeitar apenas porque veio da IA
B	Parcialmente correta	Reconhecer omissão relevante sem descartar todo o conteúdo
C	Erro conceitual	Mobilizar conhecimento disciplinar
D	Erro procedimental	Auditar cálculo, unidade ou sequência de passos
E	Conclusão correta com justificativa falha	Separar resultado de validade do raciocínio
F	Conclusão incorreta com raciocínio plausível	Resistir à persuasão superficial e localizar a ruptura lógica

5.6.5. Delimitação Interpretativa do Instrumento

O instrumento foi utilizado como recurso de investigação pedagógica interdisciplinar. A interpretação concentrou-se nas frequências de classificação dos itens e nas respostas de percepção disponíveis na tabulação consolidada. As ações abertas do V5 foram mantidas como componentes do instrumento, mas não foram convertidas em resultados por ausência dos registros individuais.

A organização do instrumento permite que novas aplicações aprofundem a análise das respostas abertas, a concordância entre avaliadores e o comportamento das diferentes dimensões do IAC-IA em outros grupos.

Tabela 3. Matriz de organização dos itens por área e tipo de situação-problema.

Área	Itens	Distribuição prevista	Controle correto
Química	6	2 conceituais; 2 quantitativos/procedimentais; 2 de crítica da IAG	Q6
Física	6	2 conceituais; 2 quantitativos/procedimentais; 2 de crítica da IAG	Q12
Biologia	6	2 conceituais; 2 quantitativos/procedimentais; 2 de crítica da IAG	Q18
Matemática	6	2 conceituais; 2 quantitativos/procedimentais; 2 de crítica da IAG	Q24

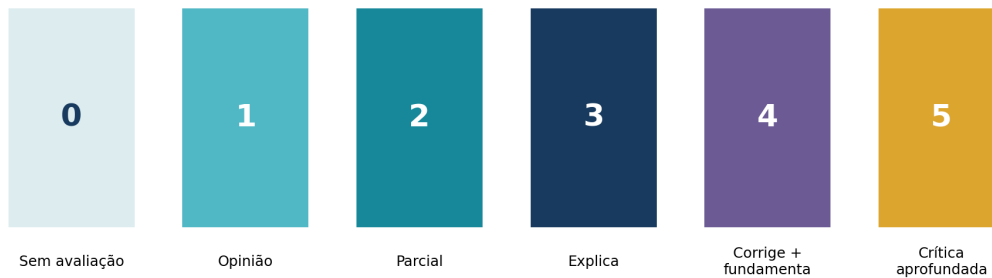
5.7. Critérios de Análise e IAC-IA

Foi definida uma escala de seis níveis, de 0 a 5, para orientar futuras codificações das respostas abertas. O nível 0 corresponde à ausência de avaliação ou à simples reprodução da resposta; o nível 1, ao julgamento sem fundamento; o nível 2, à identificação parcial; o nível 3, à localização e explicação adequadas; o nível 4, à correção fundamentada; e o nível 5, à análise de condições, alcance, incertezas ou alternativas. Como as respostas individuais não estavam disponíveis na base agregada desta versão, os níveis não foram atribuídos aos estudantes e não sustentam os resultados quantitativos apresentados.

Figura 7. Progressão qualitativa da escala de avaliação.

Rubrica de progressão da avaliação crítica

Da aceitação/rejeição intuitiva à análise fundamentada de validade e limitações



Fonte: elaboração própria.

Nível	Descritor operacional	Evidência típica
0	Não avalia ou apenas reproduz	Copia a resposta ou deixa sem julgamento
1	Julga sem fundamento	“Está errado” sem razão disciplinar
2	Identifica parcialmente	Percebe problema, mas explica de modo incompleto
3	Identifica e explica	Localiza o erro e mobiliza conceito adequado
4	Corrige e fundamenta	Reconstrói a resposta com evidência/cálculo/relação válida
5	Avalia criticamente limites	Além de corrigir, discute condições, alcance, incerteza ou alternativa

5.7.1. Dimensões do IAC-IA

Denomina-se IAC-IA, neste trabalho, o quadro de Avaliação Crítica de Respostas de Inteligência Artificial. A estrutura contempla Detecção, Localização, Verificação, Fundamentação, Correção e

Argumentação. Nesta versão, essas dimensões organizam conceitualmente o instrumento, mas não foram codificadas separadamente porque a base disponível contém resultados agregados da classificação de validade e da escala de percepção. Assim, as conclusões empíricas não atribuem níveis de desempenho no IAC-IA aos participantes.

Dimensão	Pergunta analítica	Exemplo de evidência
D: Detecção	O estudante percebe que há problema?	Classificação adequada da validade
L: Localização	Consegue apontar onde?	Indica passo, conceito ou inferência problemática
V: Verificação	Como testa a resposta?	Refaz cálculo, confronta evidência, usa contraexemplo
F: Fundamentação	Que conhecimento sustenta o julgamento?	Lei, conceito, modelo ou propriedade pertinente
C: Correção	Consegue reconstruir?	Apresenta solução ou explicação adequada
A: Argumentação	Liga evidência e conclusão?	Justificativa coerente e explicitada

5.8. Plano de Análise

O plano de análise relacionou objetivos, indicadores, tratamento descritivo e forma de apresentação. Os 24 itens foram mantidos com sua identificação original, e os quatro controles corretos foram examinados separadamente.

A análise quantitativa utiliza exclusivamente a classificação esperada registrada no Item A e a escala de percepção. As respostas abertas de localização, verificação, correção e justificativa integram o desenho do instrumento, mas não estavam disponíveis em formato individual para codificação sistemática. Por essa razão, não foram produzidos escores V5/IAC-IA nem inferências sobre o percurso individual de raciocínio.

A numeração Q1 a Q24 foi preservada nas tabelas e figuras para manter a rastreabilidade entre instrumento e resultados.

5.9. Procedimentos de Análise Quantitativa

As respostas foram sintetizadas por frequências absolutas e relativas. Para cada uma das 24 situações-problema, calculou-se a proporção de estudantes que selecionou a classificação esperada no Item A. Nos itens Q6, Q12, Q18 e Q24, a resposta apresentada pela IAG era correta e funcionava como controle; nos demais, havia erro ou limitação. Também foram calculadas médias descritivas das taxas de classificação esperada por área disciplinar e médias dos dez itens da escala de percepção. Em razão da disponibilidade apenas de dados agregados, não foram conduzidos testes inferenciais, correlações em nível individual ou comparações entre séries.

Objetivo	Indicador	Tratamento descritivo	Apresentação
Caracterizar amostra	Série, idade e uso de IAG	n e %	Texto descritivo
Desempenho por item	Classificação esperada	n e % por questão	Figura 8.

Síntese por área	Taxa esperada nos 6 itens	Média descritiva	Tabela e Figura 9
Itens de controle	Q6, Q12, Q18 e Q24	n e % de reconhecimento	Texto comparativo
Percepção	10 afirmações Likert	Média e % de concordância	Figura 10.
Integração	Desempenho × percepção	Comparação descritiva	Discussão integrada

5.9.1. Critérios de Interpretação e Transparência

A interpretação priorizou magnitude e padrão das frequências, evitando transformar diferenças descritivas em evidência de efeito causal. As quatro questões de controle foram examinadas separadamente para verificar se os estudantes simplesmente rejeitavam respostas atribuídas à IAG. Percentuais foram calculados sobre N = 83 e apresentados com uma casa decimal. Diferenças pequenas entre percentuais devem ser interpretadas com cautela, sobretudo por se tratar de amostra de uma única escola.

5.10. Integração das Evidências

O instrumento solicitou, além da classificação da validade, a localização do problema, o procedimento de verificação, a correção e a justificativa. Entretanto, a análise apresentada neste artigo limita-se às classificações e percepções agregadas. As respostas abertas deverão ser preservadas e codificadas individualmente, com manual de critérios e concordância entre avaliadores, para uma avaliação empírica posterior do V5/IAC-IA.

5.10.1. Rastreabilidade Analítica

Para preservar a rastreabilidade, as interpretações apresentadas nos Resultados e na Discussão foram vinculadas exclusivamente às frequências, percentuais e médias da tabulação consolidada. Inferências sobre motivos individuais, confiança metacognitiva ou efeito da intervenção foram evitadas quando não sustentadas diretamente pelos dados disponíveis.

5.10.2. Integração Entre Desempenho e Percepção

A comparação entre desempenho e percepção foi feita em nível grupal e descritivo. Ela permite observar a distância entre a necessidade declarada de cautela e as taxas observadas de classificação esperada, mas não permite estabelecer associação individual entre essas variáveis.

6. RESULTADOS

Para facilitar a leitura, os resultados são apresentados em correspondência com as perguntas de pesquisa: caracterização da amostra e uso de IAG; desempenho nos 24 itens; comparação descritiva entre áreas; reconhecimento das respostas corretas de controle; e percepção dos estudantes. A relação entre percepção e desempenho é discutida apenas em nível agregado.

6.1. Caracterização da Amostra e Uso de IAG

A distribuição por série foi praticamente equilibrada: 28 estudantes do 1º ano (33,7%), 28 do 2º ano (33,7%) e 27 do 3º ano (32,5%). A faixa de 16-17 anos concentrou 56,6% da amostra, seguida de 14-15 anos (31,3%) e 18 anos ou mais (12,1%). Quanto à frequência de uso de IAG, 37,3% declararam utilizar às vezes, 30,1% frequentemente e 12,1% quase todos os dias. Somados, 79,5% relataram uso pelo menos

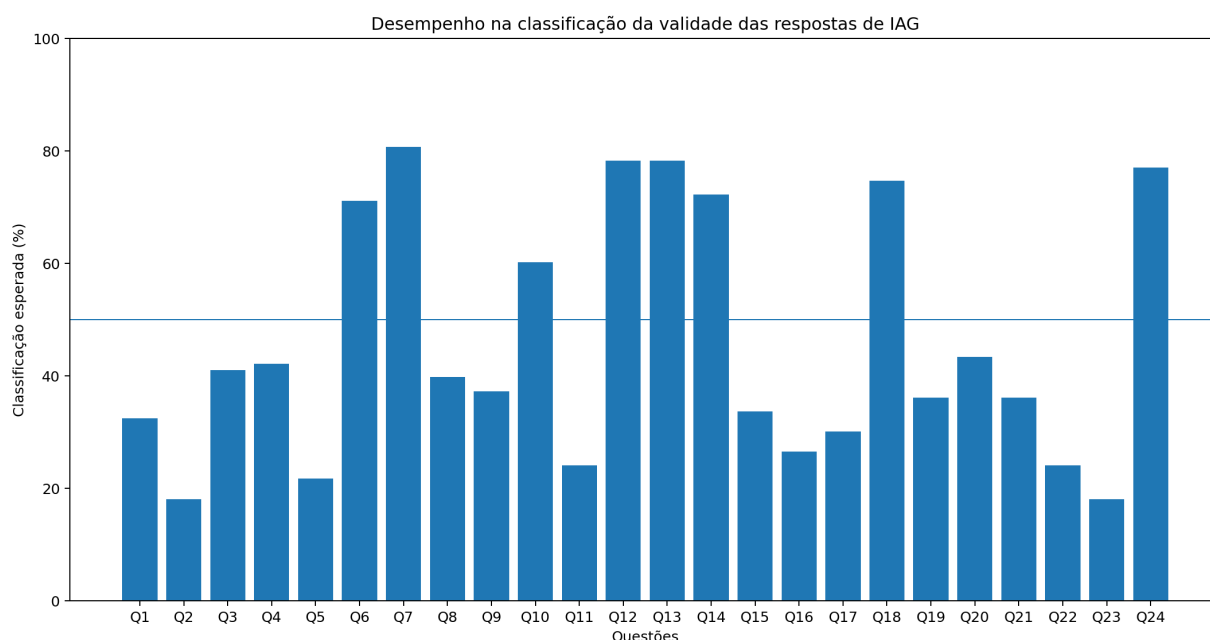
ocasional, enquanto 4,8% afirmaram nunca utilizar. Esse perfil indica familiaridade relevante com ferramentas generativas, mas não permite inferir qualidade, finalidade ou profundidade de uso além das categorias registradas.

6.2. Desempenho na Classificação da Validade

A taxa de classificação esperada variou de 18,1% a 80,7%, revelando forte heterogeneidade entre os itens. Os maiores percentuais ocorreram na questão Q7, sobre velocidade média (80,7%); na Q13, sobre fotossíntese (78,3%); na Q12, que funcionou como controle sobre movimento retilíneo uniforme (78,3%); na Q24, sobre independência de lançamentos (77,1%); na Q18, sobre seleção natural (74,7%); e na Q14, sobre antibióticos e vírus (72,3%). Os menores percentuais foram observados na Q2, sobre pH e neutralidade (18,1%); na Q23, sobre determinação de função (18,1%); na Q5, sobre propriedades e ponto de ebulição (21,7%); na Q11, sobre velocidade zero e aceleração (24,1%); e na Q22, sobre outlier e média (24,1%). Os quatro itens corretos de controle foram reconhecidos por 71,1% a 78,3% dos participantes. Esse padrão é metodologicamente relevante porque reduz a hipótese de uma estratégia indiscriminada de marcar todas as respostas da IAG como incorretas.

Área	Média descritiva dos 6 itens	Menor taxa	Maior taxa
Química	37,8%	18,1%	71,1%
Física	53,4%	24,1%	80,7%
Biologia	52,6%	26,5%	78,3%
Matemática	39,2%	18,1%	77,1%

Figura 8. Percentual de participantes que selecionaram a classificação esperada em cada questão.

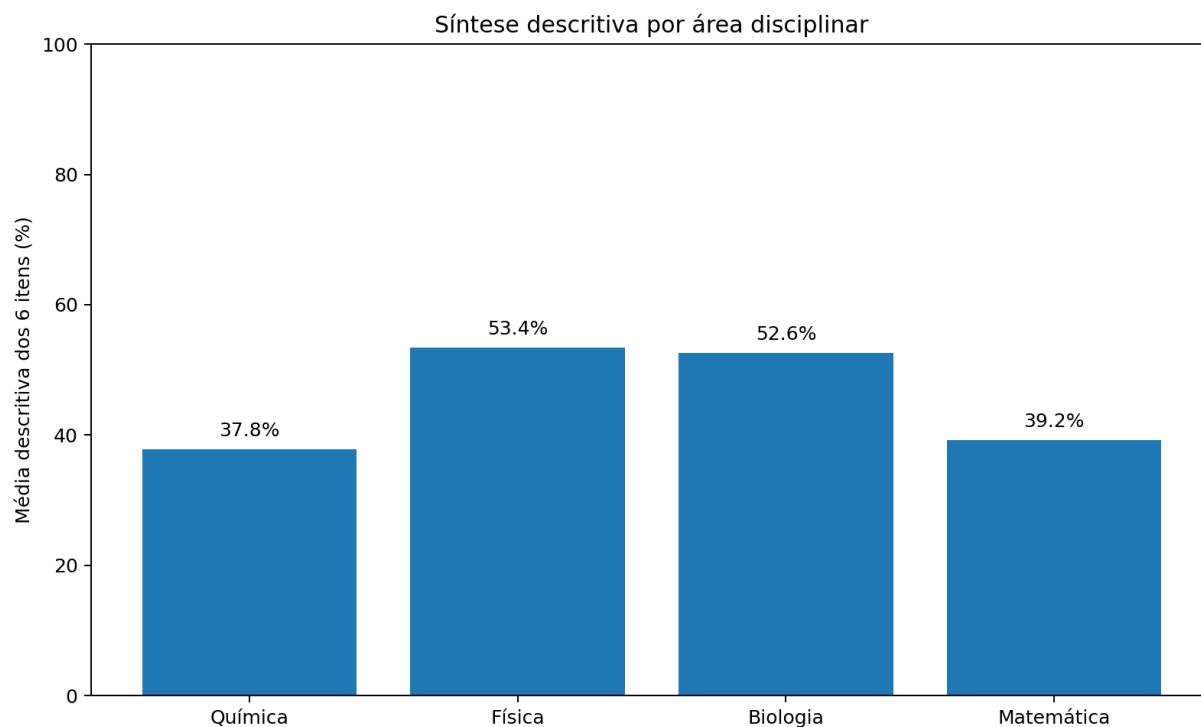


Fonte: dados da pesquisa (N = 83).

6.3. Comparação Descritiva Entre as Áreas

A média simples das taxas de classificação esperada nos seis itens de cada área foi de 37,8% em Química, 53,4% em Física, 52,6% em Biologia e 39,2% em Matemática. Esses valores são apenas sínteses descritivas do conjunto de itens aplicado. Como não foi demonstrada equivalência de dificuldade entre as questões, eles não constituem uma escala padronizada nem permitem estabelecer um ranking de proficiência entre disciplinas.

Figura 9. Média descritiva da taxa de classificação esperada nos seis itens de cada área.

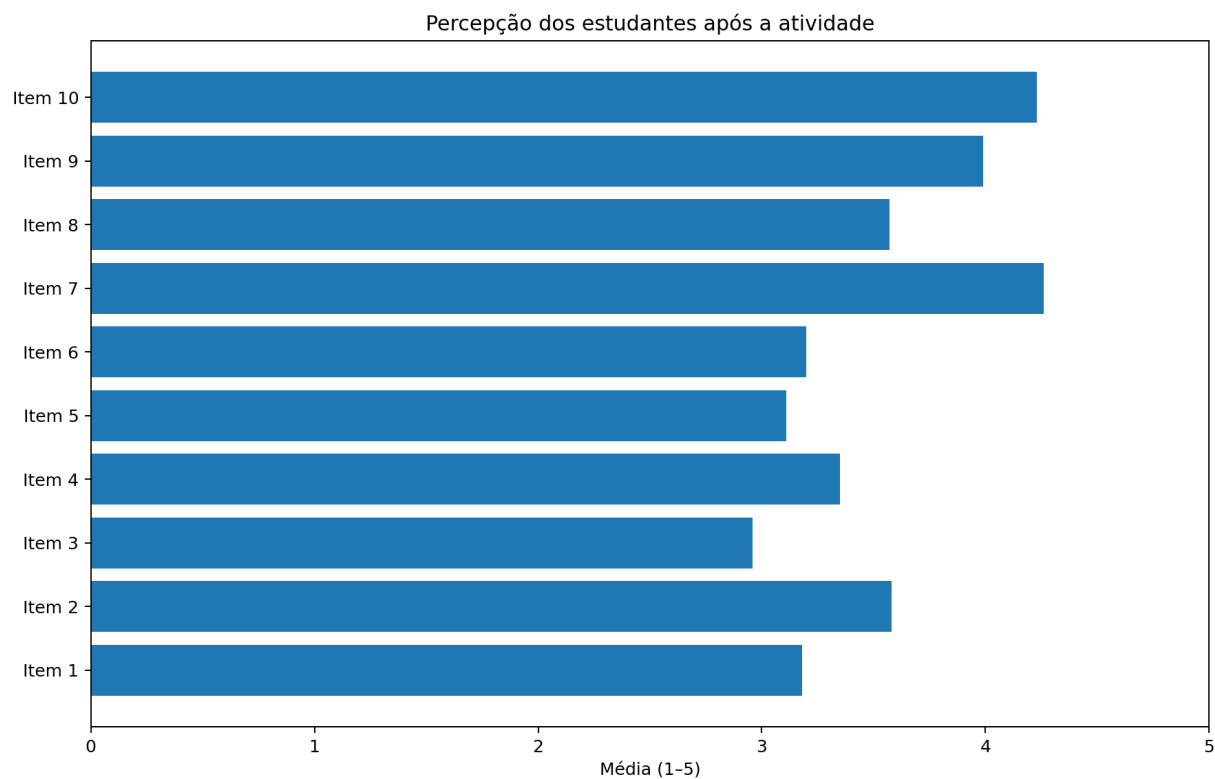


Fonte: dados da pesquisa (N = 83).

6.4. Percepção dos Estudantes Após a Atividade

Na escala de percepção, os maiores escores médios foram observados em “resposta bem escrita não é necessariamente correta” (4,26/5) e “pretendo verificar com mais cuidado antes de usar” (4,23/5). No primeiro item, 81,9% marcaram 4 ou 5; no segundo, 80,7%. A afirmação de que comparar respostas com conhecimentos ajuda a aprender apresentou média 3,99/5, com 73,5% de concordância parcial ou total. Em contraste, “consigo identificar exatamente onde está o erro” apresentou a menor média (2,96/5), e “consigo corrigir resposta inadequada de IAG” alcançou 3,11/5. A combinação desses resultados revela uma tensão pedagógica importante: a maioria reconhece a necessidade de cautela, mas a autopercepção sobre localizar e corrigir erros é mais moderada.

Figura 10. Médias das afirmações de percepção após a atividade (escala 1-5).



Fonte: dados da pesquisa (N = 83).

6.5. Síntese Integrada dos Resultados

A leitura conjunta mostra três padrões. Primeiro, a familiaridade com IAG é alta na amostra, pois 79,5% relatam uso ao menos ocasional. Segundo, familiaridade não se traduz em desempenho uniforme: a classificação esperada supera 70% em vários itens diretos e de controle, mas fica próxima ou abaixo de 25% em problemas que exigem reconhecer condições conceituais, limites de generalização ou relações menos evidentes. Terceiro, a percepção pós-atividade é mais favorável à postura de verificação do que à autoconfiança em localizar e corrigir erros. Em conjunto, os dados sustentam a distinção entre consciência crítica sobre a tecnologia e competência disciplinar para auditar respostas concretas.

7. DISCUSSÃO

7.1. Heterogeneidade do Julgamento Crítico e Natureza do Erro

Os resultados evidenciam que a avaliação crítica de respostas produzidas por IAG não se comporta, nesta amostra, como uma competência unitária que possa ser resumida apenas pela oposição entre “saber” e “não saber” avaliar a tecnologia. A variação observada entre as 24 situações foi ampla: a classificação esperada alcançou 80,7% em Q7 (velocidade média), 78,3% em Q13 (fotossíntese) e 72,3% em Q14 (antibióticos e vírus), enquanto caiu para 18,1% em Q2 (pH e neutralidade) e Q23 (determinação de função), 21,7% em Q5 (propriedades e ponto de ebulição) e 24,1% em Q11 (velocidade nula e aceleração). Essa dispersão é substantivamente mais informativa do que uma média global, porque mostra que o julgamento depende do tipo de conhecimento mobilizado, da transparência do erro e da possibilidade de submetê-lo a uma verificação operacional.

Nos itens de melhor desempenho, o estudante dispõe de critérios relativamente diretos para testar a afirmação: refazer uma razão entre grandezas, recuperar um conceito escolar bastante estabilizado ou confrontar uma proposição com uma regra conhecida. Nos itens de menor desempenho, a resposta problemática conserva elementos de plausibilidade. O erro não se apresenta como absurdo evidente; ele se aproxima de simplificações frequentemente encontradas no aprendizado escolar, como associar neutralidade invariavelmente a pH 7, inferir ponto de ebulição apenas a partir da massa molar ou generalizar uma função a partir de evidência insuficiente. Nesses casos, avaliar criticamente a IAG exige mais do que desconfiança: exige conhecimento disciplinar capaz de produzir um critério de refutação.

Esse padrão converge conceitualmente com trabalhos que defendem que a alfabetização em IA no ensino de Ciências deve incluir avaliação das saídas, reflexão crítica e compreensão das

limitações dos modelos, e não apenas habilidade operacional de formular comandos. No ensino de Química, por exemplo, Young, Dawood e Lewis (2024) exploram a reflexão crítica sobre respostas de chatbots como componente da alfabetização em IA. Revisões recentes em educação STEM também têm destacado que os benefícios educacionais da IAG dependem da forma como a tecnologia é integrada às práticas de investigação e validação, e não simplesmente de sua presença na atividade (ERÜMIT; SARIALIOĞLU, 2025; JIANG et al., 2026; NIRI et al., 2026). Os dados deste estudo acrescentam a esse debate uma evidência situada: a vulnerabilidade não é uniforme entre conteúdos, mesmo quando o formato da tarefa permanece semelhante.

7.2. Respostas Corretas de Controle e Viés de Caça Ao Erro

A inclusão de quatro itens de controle com respostas corretas (Q6, Q12, Q18 e Q24) foi importante para a análise dos resultados. Nesses itens, as taxas de acerto variaram entre 71,1% e 78,3%, mostrando que boa parte dos estudantes conseguiu reconhecer quando a resposta apresentada pela IAG estava adequada. Esse dado é relevante porque indica que os participantes não partiram simplesmente da ideia de que todas as respostas geradas pela ferramenta continham algum tipo de erro. Por isso, a presença de questões corretas e incorretas no mesmo instrumento contribuiu para uma avaliação mais equilibrada. Se todas as situações apresentassem erros, os estudantes poderiam apenas se acostumar a procurar falhas em cada resposta, sem necessariamente analisar o conteúdo e verificar se as informações estavam realmente corretas.

Do ponto de vista pedagógico, não se espera que o estudante simplesmente aceite ou rejeite as respostas produzidas pela IAG. O

mais importante é que ele desenvolva a capacidade de analisar o conteúdo apresentado e verificar se as informações fazem sentido e estão corretas. Quando a resposta estiver adequada, o estudante deve ser capaz de reconhecer os argumentos que a sustentam. Da mesma forma, quando houver erros, é importante que consiga identificá-los, explicar o problema e propor uma correção. Assim, a atividade não busca mostrar a inteligência artificial como certa ou errada, mas estimular uma análise mais cuidadosa das informações. Nesse processo, o foco está na qualidade da resposta e nas evidências que permitem considerá-la adequada ou não.

7.3. Familiaridade de Uso e Competência de Verificação

Na caracterização da amostra, 79,5% dos estudantes relataram utilizar ferramentas de IAG pelo menos “às vezes”, sendo 30,1% usuários frequentes e 12,1% usuários quase diários. A presença cotidiana da tecnologia, portanto, já faz parte da experiência escolar de grande parcela do grupo. Contudo, essa familiaridade coexistiu com taxas baixas de classificação esperada em vários itens conceitualmente exigentes. Os resultados não permitem calcular uma associação individual entre frequência de uso e desempenho, pois a tabulação disponível é agregada, mas permitem afastar uma equivalência simplista entre exposição e domínio crítico.

Essa distinção tem implicações curriculares. Aprender a operar uma interface, produzir um prompt ou obter rapidamente uma explicação não garante que o estudante disponha de repertório para avaliar a validade da resposta recebida. A alfabetização em IA, quando articulada à educação científica e matemática, precisa ser concebida como competência híbrida: envolve conhecimentos sobre a tecnologia, mas também raciocínio disciplinar, avaliação de

evidências, controle de unidades e hipóteses, reconhecimento de limites e disposição para revisar conclusões. A literatura recente sobre IAG em STEM tem apontado precisamente para a necessidade de integrar o uso da ferramenta a processos de aprendizagem ativos e reflexivos, em vez de tratá-la como fonte autossuficiente de respostas (IYAMUREMYE et al., 2024; MARZANO, 2026; ZHAO et al., 2026).

7.4. Reconhecer Erros e Saber Corrigi-los

Os resultados da Parte C mostram que os estudantes reconhecem a importância de avaliar com cuidado as respostas produzidas pela IAG. As maiores médias foram registradas nas afirmações “resposta bem escrita não é necessariamente correta” (4,26/5) e “pretendo verificar com mais cuidado antes de usar” (4,23/5). Nos dois casos, mais de 80% dos estudantes marcaram as opções 4 ou 5. Também foi bem avaliada a afirmação de que comparar as respostas da IAG com conhecimentos já adquiridos pode contribuir para a aprendizagem, que apresentou média de 3,99/5. Esses resultados mostram que a maioria dos participantes compreende a necessidade de não aceitar automaticamente as informações fornecidas pela ferramenta.

Por outro lado, os estudantes demonstraram menos segurança quando questionados sobre sua própria capacidade de identificar e corrigir erros. A afirmação “consigo identificar exatamente onde está o erro” apresentou média de 2,96/5, enquanto “consigo corrigir resposta inadequada de IAG” alcançou 3,11/5. A diferença entre esses resultados chama atenção para um aspecto importante: reconhecer que uma resposta precisa ser verificada não significa, necessariamente, saber como fazer essa verificação. Identificar onde

está o problema, explicar o motivo do erro e apresentar uma correção exige conhecimentos e habilidades que vão além da simples percepção de que a IAG pode falhar. Nesse sentido, o protocolo V5 busca orientar o estudante durante esse processo, organizando a análise em etapas mais claras.

Os resultados também mostram que é preciso ter cuidado ao interpretar apenas o que os estudantes dizem que pretendem fazer. A intenção de verificar uma resposta antes de utilizá-la é positiva, mas não significa que o estudante já saiba realizar essa verificação de maneira adequada. Quando os dados de percepção são analisados juntamente com o desempenho nas atividades, percebe-se que os participantes reconhecem a importância de conferir as informações produzidas pela IAG, mas ainda apresentam dificuldades para identificar, explicar e corrigir alguns erros. Esse resultado aponta para a necessidade de atividades que trabalhem essas habilidades de forma prática, especialmente em situações que exigem maior atenção, interpretação e raciocínio.

7.5. Papel do Conhecimento Disciplinar na Avaliação Crítica

A avaliação de respostas produzidas pela IAG mobiliza conhecimentos específicos de cada área. Em Química, por exemplo, podem ser necessários conceitos de neutralidade, estequiometria e propriedades das substâncias. Em Física, a análise pode envolver grandezas, condições de movimento e cálculos. Em Biologia, pode exigir atenção aos mecanismos e aos limites de uma explicação. Em Matemática, pode envolver raciocínio lógico, probabilístico, estatístico ou funcional. Essa leitura é compatível com os padrões observados, mas não foi testada por análise individual das justificativas.

O desenvolvimento de uma postura crítica diante da IAG também não precisa ficar restrito às disciplinas relacionadas à tecnologia. Química, Física, Biologia e Matemática oferecem diferentes oportunidades para trabalhar a análise e a verificação das respostas geradas por essas ferramentas. Uma abordagem interdisciplinar pode contribuir para esse processo sem desconsiderar as características próprias de cada área. O estudante pode aprender a identificar possíveis erros, verificar informações, refazer cálculos, comparar explicações e justificar suas conclusões de acordo com os conhecimentos de cada disciplina.

7.6. Erro Como Recurso Didático e Avaliação Formativa

Do ponto de vista pedagógico, as respostas produzidas pela IAG podem ser utilizadas como material de aprendizagem. Em vez de serem apresentadas como respostas prontas ou gabaritos, podem servir como ponto de partida para atividades de análise e discussão, nas quais os estudantes identifiquem o que está correto, localizem problemas, refaçam cálculos e proponham correções.

Dessa forma, o erro passa a fazer parte do processo de aprendizagem e pode fornecer ao professor informações importantes sobre as dificuldades dos estudantes. Mais do que verificar se a resposta final está certa ou errada, o professor pode observar como o aluno chegou àquela conclusão e em que momento encontrou dificuldades.

Dois estudantes podem classificar corretamente uma resposta por razões distintas: um pode apenas desconfiar, enquanto outro pode localizar o problema e fundamentar uma correção. A base agregada utilizada neste artigo não permite distinguir esses percursos. O

V5/IAC-IA oferece categorias para uma análise futura das respostas abertas, mas essa interpretação dependerá da recuperação dos registros individuais, da definição de um manual de codificação e da verificação da concordância entre avaliadores.

7.7. Limitações do Estudo e Possibilidades de Novas Pesquisas

Os resultados apresentados neste estudo correspondem à realidade observada com 83 estudantes de três turmas do Ensino Médio, sendo uma turma de cada série, em uma escola pública da zona norte de Teresina. Os dados permitem compreender como esses estudantes analisaram as respostas produzidas pela IAG e quais dificuldades apareceram durante as atividades. No entanto, como a pesquisa foi realizada em uma única escola, os resultados não devem ser considerados como representativos de todos os estudantes da rede estadual ou de outras realidades educacionais.

Outro aspecto que deve ser considerado é a forma como os dados estão organizados. A base utilizada reúne os resultados de maneira geral, o que limita algumas comparações que poderiam ampliar a análise. Por exemplo, não foi possível verificar separadamente o desempenho dos estudantes do 1º, 2º e 3º anos, nem analisar com maior profundidade se a frequência de uso da IAG esteve relacionada ao desempenho nas atividades. Da mesma forma, os dados disponíveis não permitem acompanhar individualmente possíveis mudanças antes e depois da atividade ou realizar uma comparação com estudantes que não participaram da proposta. Por esse motivo, os resultados devem ser entendidos como evidências do que foi observado durante a aplicação, sem atribuir exclusivamente à atividade as diferenças encontradas.

As respostas abertas também oferecem possibilidades de análise que podem ser aprofundadas em trabalhos futuros. Uma avaliação mais detalhada dessas respostas poderá mostrar não apenas se o estudante identificou um erro, mas também como explicou o problema, quais conhecimentos utilizou e de que maneira propôs uma correção. Para isso, será importante estabelecer critérios claros de análise e avaliação das respostas.

Essas limitações não diminuem a importância dos resultados encontrados, mas ajudam a definir até onde eles podem ser interpretados. O estudo mostrou que os estudantes apresentam diferentes níveis de facilidade para avaliar respostas produzidas pela IAG e que reconhecer que a inteligência artificial pode cometer erros não significa, necessariamente, saber identificar e corrigir esses erros. Pesquisas futuras poderão ampliar a investigação para outras escolas e turmas, comparar estudantes de diferentes séries e acompanhar o desempenho antes e depois das atividades. Isso permitirá compreender melhor de que maneira estratégias como o protocolo V5 podem contribuir para o desenvolvimento de uma análise mais cuidadosa das informações produzidas pela IAG.

8. IMPLICAÇÕES PEDAGÓGICAS

A proposta não busca substituir o professor pela inteligência artificial, nem defender a proibição dessas ferramentas no ambiente escolar. A intenção é utilizá-las de uma maneira que contribua para a aprendizagem. As respostas produzidas pela IAG podem servir como material para análise, permitindo que os estudantes identifiquem informações corretas, reconheçam possíveis erros e justifiquem suas conclusões. Nesse processo, cabe ao professor selecionar exemplos adequados, propor situações que estimulem a

reflexão e acompanhar a forma como os estudantes chegam às suas respostas. Para o aluno, a atividade ajuda a compreender que uma resposta bem escrita e aparentemente convincente não deve ser considerada correta sem antes ser analisada.

O protocolo V5 pode ser utilizado em diferentes situações, como atividades em sala de aula, avaliações formativas, projetos envolvendo mais de uma disciplina e ações de formação de professores. Essas adaptações devem considerar os conhecimentos próprios de cada área, pois a análise de uma resposta depende da compreensão do conteúdo envolvido. Sem essa relação com o conhecimento trabalhado em sala de aula, a atividade pode acabar se limitando apenas a opiniões sobre o uso da tecnologia.

8.1. Relevância para o Contexto Escolar Brasileiro

Pensando especialmente na realidade da escola pública, a proposta foi organizada de forma que sua aplicação não dependa da assinatura de ferramentas comerciais de inteligência artificial por parte dos estudantes. As respostas que serão analisadas podem ser preparadas previamente pelo professor ou pesquisador e apresentadas de forma impressa ou por meio dos recursos digitais disponíveis na escola. Dessa maneira, a atividade pode ser desenvolvida mesmo em escolas com acesso limitado à internet, poucos equipamentos ou outras dificuldades de infraestrutura.

Essa possibilidade torna a proposta mais acessível a diferentes realidades escolares e permite trabalhar o uso crítico da inteligência artificial sem exigir que cada estudante tenha acesso individual a uma plataforma de IAG. Ao mesmo tempo, as diferenças de acesso às tecnologias não devem ser ignoradas. Elas também fazem parte

dessa discussão e precisam ser consideradas quando se pensa na utilização da inteligência artificial no contexto educacional brasileiro.

9. CONSIDERAÇÕES FINAIS

Este estudo descreveu como 83 estudantes de três turmas do Ensino Médio classificaram 24 respostas atribuídas à IAG em Química, Física, Biologia e Matemática. Os resultados se referem às frequências de classificação esperada e às percepções declaradas após a atividade; não medem desenvolvimento individual nem efeito causal do protocolo V5.

O desempenho variou de 18,1% a 80,7% entre os itens. As maiores taxas ocorreram em situações com critérios de verificação mais diretos e nos itens corretos de controle. As menores taxas apareceram em problemas que exigiam reconhecer condições conceituais ou relações menos evidentes. Como os itens não foram calibrados para equivalência de dificuldade, as médias por área devem ser interpretadas apenas como sínteses descritivas do instrumento aplicado.

Na escala de percepção, a maioria dos participantes reconheceu que uma resposta bem escrita não é necessariamente correta e declarou intenção de verificar as saídas de IAG com mais cuidado. Ao mesmo tempo, os próprios estudantes relataram menor segurança para localizar e corrigir erros. A comparação desses resultados é grupal e descritiva, pois a base agregada não permite relacionar percepção e desempenho no nível individual.

O protocolo V5 organizou a atividade em detectar, localizar, verificar, corrigir e justificar. Entretanto, as respostas abertas correspondentes a essas ações não estavam disponíveis em formato individual para

codificação sistemática. Por isso, o estudo não atribui escores de desempenho no V5/IAC-IA. A análise dessas dimensões dependerá da recuperação dos registros, de um manual de codificação e da concordância entre avaliadores.

A presença de Química, Física, Biologia e Matemática no mesmo instrumento permitiu observar diferenças entre situações disciplinares, mas não comprova transferência de uma estratégia crítica entre domínios. Os resultados indicam apenas que a identificação da validade variou conforme o conteúdo e o tipo de problema apresentado.

As conclusões estão limitadas aos 83 participantes de uma escola pública da zona norte de Teresina. O uso de uma turma por série, a ausência de pré-teste, pós-teste e grupo de comparação e a disponibilidade apenas de dados agregados impedem generalizações amplas e inferências sobre aprendizagem produzida pela atividade.

Dentro desses limites, o trabalho mostra que cautela declarada diante da IAG não coincide automaticamente com a identificação correta de erros em situações concretas. O instrumento oferece uma base pedagógica para atividades de verificação, mas sua capacidade de avaliar localização, fundamentação, correção e argumentação deverá ser examinada em análises posteriores das respostas abertas.

REFERÊNCIAS BIBLIOGRÁFICAS

ALFARWAN, A. A. Generative AI use in K-12 education: a systematic review. *Frontiers in Education*, v. 10, 2025. DOI: 10.3389/feduc.2025.1647573.

BOOLZEN, C.; KUHN, J.; FLEGR, S. et al. Evidence of impact and interpretational limits of generative AI in STEM education: a systematic review and meta-analysis on cognitive learning outcomes. *Artificial Intelligence Review*, 2026. DOI: 10.1007/s10462-026-11665-9.

BRASIL. Ministério da Educação. Base Nacional Comum Curricular: Ensino Médio. Brasília, DF: MEC, 2018.

ERÜMIT, A. K.; SARIALIOĞLU, R. Ö. Artificial intelligence in science and chemistry education: a systematic review. *Discover Education*, v. 4, art. 178, 2025. DOI: 10.1007/s44217-025-00622-3.

IYAMUREMYE, A. et al. Utilization of artificial intelligence and machine learning in chemistry education: a critical review. *Discover Education*, v. 3, art. 95, 2024. DOI: 10.1007/s44217-024-00197-5.

JIANG, H.; LI, Y.; CHUGH, R.; XIN, Y.; CHENG, H. The AI-powered co-inquirer: a systematic review of ChatGPT for inquiry-based learning in STEAM education. *International Journal of STEM Education*, v. 13, art. 42, 2026. DOI: 10.1186/s40594-026-00628-9.

LIU, X.; ZHONG, B. Integrating generative artificial intelligence into student learning: a systematic review from a TPACK perspective. *Educational Research Review*, v. 49, art. 100741, 2025. DOI: 10.1016/j.edurev.2025.100741.

MARZANO, D. Generative Artificial Intelligence (GAI) in Teaching and Learning Processes at the K-12 Level: A Systematic Review. *Technology, Knowledge and Learning*, v. 31, p. 789-829, 2026. DOI: 10.1007/s10758-025-09853-7.

MIAO, F.; HOLMES, W. Guidance for generative AI in education and research. Paris: UNESCO, 2023.

NIRI, G. et al. STEAM education for AI literacy: a systematic literature review. *International Journal of STEM Education*, v. 13, art. 46, 2026. DOI: 10.1186/s40594-026-00629-8.

VELDHUIS, A.; LO, P. Y.; KENNY, S.; ANTLE, A. N. Critical Artificial Intelligence literacy: a scoping review and framework synthesis. *International Journal of Child-Computer Interaction*, v. 43, art. 100708, 2025. DOI: 10.1016/j.ijcci.2024.100708.

YOUNG, J. D.; DAWOOD, L.; LEWIS, S. E. Chemistry Students' Artificial Intelligence Literacy through their Critical Reflections of Chatbot Responses. *Journal of Chemical Education*, v. 101, n. 6, p. 2466-2474, 2024. DOI: 10.1021/acs.jchemed.4c00154.

ZHAO, J.; WANG, Z.; LUO, X. et al. The influence of generative artificial intelligence on STEM education: a systematic review of recent empirical studies. *Humanities and Social Sciences Communications*, 2026. DOI: 10.1057/s41599-026-08876-4.

ZHONG, B.; LIU, X. Evaluating AI literacy of secondary students: framework and scale development. *Computers & Education*, art. 105230, 2024. DOI: 10.1016/j.compedu.2024.105230.

¹ Licenciatura Plena em Química pela Universidade Estadual do Piauí (UESPI), Teresina, Piauí, Brasil. Mestre e doutorando em Ciência e Engenharia de Materiais pela Universidade Federal do Piauí (UFPI), Teresina, Piauí, Brasil. E-mail: [acesse o artigo original para visualizar o](#)

e-mail. ORCID: <https://orcid.org/0009-0008-9112-758X>. Lattes: <https://lattes.cnpq.br/0156962783585205>

² Doutor em Educação pela Universidade Federal do Piauí. E-mail: [acesse o artigo original para visualizar o e-mail](#). LATTES: <http://lattes.cnpq.br/3777287905500590>. ORCID: <https://orcid.org/0000-0001-6914-9012>

³ Licenciatura Plena em Química pela Universidade Estadual do Piauí (UESPI), Teresina, Piauí, Brasil. Mestre em Química pela Universidade Estadual do Piauí (UESPI), Teresina, Piauí, Brasil. E-mail: [acesse o artigo original para visualizar o e-mail](#). ORCID: <https://orcid.org/0000-0003-0565-3302>

⁴ Doutor em Ciências Biológicas (Zoologia) pela UFPB. Universidade Federal do Piauí (UFPI), Departamento de Biologia, Laboratório de Etnobiologia e Sustentabilidade (LES). ORCID: <https://orcid.org/0000-0001-7508-9175>. Lattes: <http://lattes.cnpq.br/6517728431299455>. E-mail: [acesse o artigo original para visualizar o e-mail](#)

⁵ Mestre em Ensino de Matemática pela Universidade do Estado do Pará (UEPA). E-mail: [acesse o artigo original para visualizar o e-mail](#). ORCID: <https://orcid.org/0000-0001-7102-0322>. Lattes: <https://lattes.cnpq.br/0262016537459226>

⁶ Mestranda em Ensino de Ciências da Natureza pela Universidade Federal do Piauí (UFPI). E-mail: [acesse o artigo original para visualizar o e-mail](#). Lattes: <http://lattes.cnpq.br/6230478417982638>

⁷ Mestrado em física da matéria condensada. Universidade federal do Piauí E-mail: [acesse o artigo original para visualizar o e-mail](#)

⁸ Mestrado em Ensino de Biologia pela Universidade Estadual do Piauí (UESPI). E-mail: [acesse o artigo original para visualizar o e-mail](#)

⁹ Licenciatura Plena em Matemática, pela Universidade Federal do Piauí-UFPI. Licenciatura Plena em Pedagogia, pelo Instituto de Ensino Superior Múltiplo-IESM. Pós-Graduado em: Docência para o Ensino Superior, pelo Instituto de Ensino Superior Múltiplo-IESM; Gestão e Supervisão Escolar, pelo Instituto de Ensino Superior Múltiplo-IESM; Língua Brasileira de Sinais-LIBRAS, pelo Instituto de Ensino Superior Múltiplo-IESM.

¹⁰ Especialista em Metodologia de Ensino pelo Instituto de Ensino Superior Múltiplo (IESM). E-mail: [acesse o artigo original para visualizar o e-mail](#). Lattes: 1539981073873187

¹¹ Licenciatura em Química. Instituto Federal do Piauí. E-mail: [acesse o artigo original para visualizar o e-mail](#). Lattes: <http://lattes.cnpq.br/9866158542688932>

¹² ESPECIALIZAÇÃO EM PROGRAMAÇÃO PARA DISPOSITIVOS MÓVEIS. Instituto de Tecnologia do Piauí (IFPI). Lattes: <http://lattes.cnpq.br/3166474684334656>

¹³ Graduado em Química pelo Instituto Federal de Educação, Ciência e Tecnologia do Piauí (IFPI). E-mail: [acesse o artigo original para visualizar o e-mail](#). Lattes: <http://lattes.cnpq.br/8009915688479956>. ORCID: <https://orcid.org/0009-0001-5673-6549>

¹⁴ Especialização em Banco de Dados pela Faculdade de Tecnologia do Piauí (FATEPI). E-mail: [acesse o artigo original para visualizar o e-mail](#)

mail. Lattes: <http://lattes.cnpq.br/7356682585955101>. ORCID:
<https://orcid.org/0009-0002-7139-0539>

¹⁵ Licenciado em Física pela Universidade Federal do Piauí (UFPI).
Mestrado em Engenharia de Materiais pelo Instituto Federal do Piauí
(IFPI). Lattes: <http://lattes.cnpq.br/450028242947274>. ORCID:
<https://orcid.org/0009-0003-1823-9802>

¹⁶ Mestrado em Engenharia de Materiais. Instituição: Secretaria de
Estado da Educação (SEDUC). E-mail: [acesse o artigo original para
visualizar o e-mail](#). ORCID: <https://orcid.org/0009-0009-1899-9573>.
Lattes: <http://lattes.cnpq.br/7626607416007220>