Avaliação Automatizada de Relatórios Experimentais de Física com Gemini AI Studio: Um Relato de Experiência e Análise Comparativa

Tábata Caroline Gonçalves Mendes[1], Kenya Aparecida Alves[2]

[1]Instituto Federal de São Paulo Campus São José dos Campos (IFSP-SJC)
São José dos Campos – SP – Brazil

[2]Instituto Federal de São Paulo Campus São José dos Campos (IFSP-SJC)
São José dos Campos – SP – Brazil.

tabata.caroline@aluno.ifsp.edu.br, kenya.aparecida@ifsp.edu.br

Abstract. *This experience report compares human evaluation of Physics experimental reports with automated assessment using Gemini AI Studio. Analyzing four works from Engineering students, the main objective was to investigate the viability and precision of artificial intelligence as a support tool for academic evaluation. The AI proved accurate in detecting structural and conceptual issues. Structured prompts were developed using prompt engineering techniques to guide Gemini AI Studio in evaluating the reports. The results revealed that while AI demonstrates capability in identifying structural and formal issues, human evaluation proved superior in contextual analysis and conceptual interpretation. The significant discrepancies found in final grades, with differences of up to 80% between evaluators, evidence important limitations of the automated approach, indicating that human pedagogical supervision remains indispensable for adequate formative assessment.*

Resumo. *Este relato de experiência compara a avaliação humana de relatórios experimentais de Física com uma avaliação automatizada utilizando Gemini AI Studio. Analisando quatro trabalhos de estudantes de Engenharia, o objetivo principal foi investigar a viabilidade e precisão da inteligência artificial como ferramenta de apoio à avaliação acadêmica. A IA mostrou-se precisa na detecção de problemas estruturais e conceituais. Foram desenvolvidos prompts estruturados utilizando técnicas de engenharia de prompt para orientar o Gemini AI Studio na avaliação dos relatórios. Os resultados revelaram que, embora a IA demonstre capacidade para identificar problemas estruturais e formais, a avaliação humana mostrou-se superior em análises contextuais e na interpretação conceitual. As significativas divergências encontradas nas notas finais, diferenças de até 80% entre avaliadores, evidenciam limitações importantes da abordagem automatizada, indicando que a supervisão pedagógica humana permanece indispensável para uma avaliação formativa adequada.*

1. Introdução

A avaliação de relatórios técnico-científicos assume um papel importante na formação de estudantes de disciplinas experimentais em cursos de Ciências da Natureza (Química, Biologia, Física) e, particularmente, na Engenharia. Os relatórios são muito

DOI: 10.5753/wie.2025.13512

mais do que um simples registro de procedimentos; funcionam como verdadeiras ferramentas de aprendizagem, onde os estudantes desenvolvem a capacidade de se comunicar como cientistas, analisar dados com um olhar crítico e construir argumentos sólidos [Lima *et al*., 2017]. Ao elaborar um relatório, o estudante é desafiado a conectar a teoria com a prática, transformando um experimento em uma narrativa coerente e bem fundamentada, uma competência essencial para sua futura carreira [Araújo; Abib, 2003].

Apesar de sua importância formativa, o processo avaliativo desses relatórios impõe desafios significativos aos docentes. A correção detalhada consome um tempo enorme do docente, pois exige uma análise minuciosa de múltiplos critérios, como, clareza metodológica, precisão conceitual, adequação às normas e qualidade da análise de resultados. Em turmas grandes, o volume de trabalho pode sobrecarregar o docente, tornando difícil manter um padrão de avaliação justo e oferecer um feedback individualizado que realmente ajude o estudante a melhorar [Barreto; Costa, 2017]. Além disso, a subjetividade é um fator inevitável; mesmo com guias de correção claros (rubricas), a interpretação do Avaliador Humano sempre terá um papel, o que pode levar a variações na avaliação [Nunes *et al*., 2021].

Nesse contexto, a Inteligência Artificial (IA) generativa, viabilizada por Grandes Modelos de Linguagem (LLMs), como o Gemini, apresenta-se como uma poderosa ferramenta que pode transformar a educação. Trabalhos como o de Mok *et al.* (2024), que testaram a IA para corrigir trabalhos de Física, mostram que o interesse em usar essa tecnologia para otimizar a avaliação está crescendo rapidamente [Mok *et. al.,* 2024]. A promessa é que a IA possa não apenas agilizar o processo, mas também trazer mais consistência para a correção de aspectos objetivos.

Mas como fazer uma IA avaliar algo tão complexo quanto um relatório científico? A resposta está na Engenharia de *Prompt*, que consiste na formulação criteriosa de instruções dirigidas ao modelo IA. Entre as técnicas mais eficazes, destaca-se o *Chain-of-Thought Prompting* (Cadeia de Raciocínio), cuja eficácia foi demonstrada por Wei *et al.* (2022), ao evidenciar que instruções que solicitam a explicitação do raciocínio da IA etapa por etapa aumentam substancialmente a precisão das respostas. Em vez de apenas dar a resposta final, a IA "mostrava seu trabalho". Embora o foco deles fosse a resolução de exercícios, o princípio é diretamente aplicável à avaliação escrita: podemos pedir à IA que não apenas dê uma nota, mas que justifique cada ponto da sua análise, simulando o raciocínio detalhado de um avaliador humano e tornando seu processo mais transparente e confiável [Wei *et. al*., 2022].

Apesar desses avanços técnicos, ainda existe uma lacuna importante na pesquisa: faltam estudos que realizem comparações sistemáticas entre as avaliações conduzidas pela IA com a avaliação de um Avaliador Humano especialista, especialmente no contexto do ensino superior brasileiro e em relatórios que exigem análise de nuances metodológicas. É exatamente essa lacuna que este trabalho busca investigar. Apresentamos aqui um relato de experiência que documenta uma análise comparativa entre a avaliação tradicional, feita por um docente, e uma avaliação automatizada, conduzida com o *Gemini AI Studio*.

Nossa investigação se concentrou em quatro relatórios de Física Experimental, e nosso objetivo foi investigar a viabilidade e a precisão da IA como uma ferramenta de apoio. Ao comparar os dois métodos, pudemos identificar onde a análise da IA converge com a do Avaliador Humano e onde ela diverge, revelando tanto seu potencial quanto

suas atuais limitações. Com base nos resultados, este trabalho pretende contribuir com reflexões práticas para docentes e pesquisadores interessados em incorporar ferramentas baseadas em IA ao processo educativo, promovendo sua adoção responsável e crítica, sem abdicar do papel essencial da sensibilidade pedagógica do olhar humano.

2. Metodologia

A motivação desta pesquisa reside na compreensão da avaliação não como uma etapa meramente classificatória, mas como um componente essencial do processo pedagógico. Partimos do entendimento, alinhado aos pressupostos da Teoria da Aprendizagem Significativa de David Ausubel, de que o feedback detalhado contribui para que o estudante possa reorganizar suas ideias e construir conhecimentos mais elaborados, ancorando novas informações à sua estrutura cognitiva pré-existente.

Nosso objetivo principal foi garantir a replicabilidade e a transparência da análise comparativa entre a avaliação humana e a automatizada, abordando os desafios inerentes à correção de relatórios técnico-científicos.

A metodologia deste estudo foi, portanto, estruturada como um processo iterativo de desenvolvimento e validação, fundamentado nesses princípios pedagógicos e executado com rigor técnico para garantir a replicabilidade e a transparência da análise comparativa.

O estudo foi desenvolvido no contexto da disciplina de Física de Laboratório de Mecânica e Ondas, ministrada no terceiro semestre do curso de Engenharia de Produção de uma instituição pública federal. Os relatórios analisados foram os primeiros produzidos pelos discentes dentro de uma perspectiva técnico-científica formal, constituindo-se como um momento pedagógico de transição entre a linguagem cotidiana e a linguagem acadêmica especializada. Com o intuito de orientar essa etapa inicial, a docente forneceu previamente um modelo de estrutura de relatório e uma rubrica de avaliação detalhada, com ênfase nas normas da ABNT (como NBR 10719 e NBR 6023), para explicitar os critérios de qualidade esperados.

O desenho do estudo previu uma comparação controlada. Foram selecionados quatro relatórios completos, mantidos em sua forma original, sem correções ou anotações, os quais foram submetidos ao Avaliador IA. Em paralelo, foi utilizada a versão dos mesmos relatórios já corrigida pelo Avaliador Humano no ano de 2022, contendo seus comentários qualitativos e a nota atribuída. Esta versão corrigida funcionou como nosso padrão-ouro (*gold standard*), permitindo que a avaliação automatizada fosse conduzida em uma condição de correção cega.

A ferramenta de Inteligência Artificial selecionada para a avaliação automatizada foi *o Gemini AI Studio*[1*], utilizando sua modalidade de uso gratuito, com o modelo avançado *Gemini 1.5 Pro*. Nossa escolha dessa plataforma foi estratégica, fundamentada em duas de suas características mais notáveis, cruciais para a tarefa de avaliação: sua excepcional capacidade de Processamento de Linguagem Natural (PLN) e, decisivamente, sua ampla janela de contexto (1.048.576 tokens). Essa vasta "memória" permitiu que a IA analisasse integralmente relatórios longos (o maior deles, por exemplo, consumiu 43.226 tokens), sem a necessidade de cortes ou resumos prévios. Igualmente

[1] GOOGLE. Gemini AI Studio. Mountain View: Google, 2024. Disponível em: https://aistudio.google.com/. Acesso em: 18 jun. 2025.

importante, garantiu que o modelo pudesse aderir a instruções complexas e detalhadas de nosso *prompt* de avaliação.

Para assegurar a replicabilidade do experimento, o modelo foi configurado com parâmetros específicos combinando escolhas deliberadas e padrões da plataforma. A Temperatura controla o "grau de criatividade" das respostas: valores baixos (próximo de zero) geram respostas previsíveis, enquanto valores altos (próximo de um) incentivam diversidade. Configuramos em 1.0 (máximo) para testar os limites da IA em formular feedback construtivo e menos mecânico, evitando respostas robóticas dentro dos limites estruturais do *prompt*. O *Top-P* refina a seleção lexical definindo limite de probabilidade - com valor 0.95, o modelo considera apenas palavras que somam 95% de probabilidade cumulativa, ignorando os 5% menos prováveis. Esta configuração equilibrou a criatividade da alta temperatura com respostas semanticamente coerentes, filtrando opções improváveis. O Comprimento da Saída foi configurado em 65.536 tokens para evitar interrupções da avaliação, permitindo análise completa mesmo nos relatórios mais extensos.

Estabelecemos restrições específicas nos modos de processamento visando integridade metodológica. O *Thinking Mode*, que permite pré-processamento interno antes da resposta final, foi desabilitado porque nossa estratégia já incorporava *Chain-of-Thought Prompting* explicitamente. Preferimos forçar a IA a "pensar em voz alta", documentando suas oito etapas de avaliação na resposta final para total transparência do raciocínio. Todas as ferramentas externas foram desabilitadas: *Structured Output* (formatos específicos como JSON), *Code Execution* (execução de códigos), *Function Calling (API*s externas), *Grounding with Google Search* (buscas na internet) e *URL Context* (análise de páginas web). Esta desativação garantiu que a avaliação se baseasse estritamente no conteúdo dos relatórios e instruções do *prompt*, criando ambiente isolado que mimetiza as condições do avaliador humano para comparação metodológica justa.

Entre as limitações metodológicas identificadas, não foram testadas diferentes configurações dos paramétricas, utilizando-se as configurações padrão da plataforma. Embora as escolhas de desabilitação de ferramentas externas tenham sido metodologicamente justificadas, a exploração sistemática de diferentes configurações paramétricas permanece como oportunidade para otimização futura do processo.

Cabe destacar que este estudo não envolveu desenvolvimento de código programático tradicional, baseando-se exclusivamente em técnicas de engenharia de *prompt* - isto é, na elaboração criteriosa de instruções estruturadas em linguagem natural para orientar o comportamento da IA.

Para traduzir nossos princípios pedagógicos em uma ferramenta de avaliação prática, nosso foco se voltou para o desenvolvimento do instrumento central deste estudo: o *prompt* de avaliação. A construção deste instrumento não foi um evento único, mas uma jornada de engenharia iterativa, marcada por uma evolução significativa que partiu de uma concepção inicial promissora para um modelo final robusto e detalhado.

Nossa abordagem inicial, na Fase 1, foi de natureza meta-cognitiva. Em vez de construir o *prompt* do zero, fornecemos ao próprio *Gemini AI Studio* uma lista detalhada com todos os critérios de avaliação da disciplina — abrangendo desde a formatação da capa até o rigor científico da conclusão — e solicitamos que ele mesmo gerasse uma estrutura para avaliar os relatórios. A IA respondeu de forma impressionante, criando um *template* de avaliação bem organizado, que dividia a análise em seções lógicas como

"Análise Geral", "Análise de Conteúdo" e "Checklist de Penalidades".

No entanto, a Fase 2, de teste e refinamento, rapidamente expôs as limitações dessa abordagem. Ao aplicar este *prompt* inicial em um teste piloto com quatro relatórios, os resultados foram desapontadores. A avaliação gerada, embora estruturada, mostrou-se superficial, e as notas finais apresentaram uma alta divergência em relação às atribuídas pelo avaliador humano, com um erro relativo médio de 150,6%. Essa falha inicial tornou evidente que apenas fornecer os critérios não era suficiente; era preciso guiar o "raciocínio" da IA de forma muito mais explícita, controlando não só o que avaliar, mas como avaliar.

Essa necessidade de refinamento nos levou à Fase 3, de otimização, onde adotamos uma abordagem mais robusta para construir o instrumento final. Para resolver a superficialidade e a falta de transparência, aplicamos um conjunto de técnicas avançadas de engenharia de *prompt*, sendo elas: o *Role-Based Prompting* (Definição de Papel): Para ajustar o tom e a perspectiva da avaliação, instruímos a IA a assumir a persona de um "*professor experiente de Física Experimental... com expertise em Metodologia Científica, Normas ABNT e Pedagogia Universitária*". O *Chain-of-Thought Prompting* (Cadeia de Pensamento): Para combater a arbitrariedade das notas, impusemos uma sequência de raciocínio obrigatória em 8 etapas. Isso forçou a IA a detalhar sua análise passo a passo, tornando seu processo transparente e evitando que a nota final fosse um "chute" monolítico. A *Structured Criteria & Output Formatting*: Para garantir consistência, fornecemos um modelo de saída rígido, com penalidades e pontuações específicas para cada erro, guiando tanto a análise quanto a apresentação do resultado. E o *Few-Shot Learning* (Microexemplos): Embora não tenhamos fornecido relatórios completos como exemplos, inserimos microexemplos de feedback esperado diretamente no *prompt* (ex: [*A contextualização inicial... foi bem-feita.*]), guiando a IA sobre o formato e o tipo de comentário esperado.

O processo iterativo culminou no desenvolvimento do instrumento de avaliação final, um *prompt* estruturado em oito etapas que orientou toda a avaliação automatizada. O trecho abaixo exemplifica a estrutura deste instrumento:

"DEFINIÇÃO DE PAPEL
Haja como um professor experiente de Física Experimental do curso superior de Engenharia de Produção (1º semestre), com expertise em Metodologia Científica, Normas ABNT e Pedagogia Universitária. Sua missão é fornecer uma avaliação rigorosa, criteriosa, transparente e altamente construtiva de relatórios técnicos científicos.
PROCESSO DE AVALIAÇÃO EM 8 ETAPAS OBRIGATÓRIAS
Execute esta sequência de forma inflexível. A saída deve ser um documento completo e estruturado, mostrando todo o processo de avaliação.
NOTA: introdução tem nota máxima de 2,0 a fundamentação teórica nota máxima de 2,0, procedimento experimental nota máxima de 1,0, resultados e discussões nota máxima de 3,0, e conclusão 2,0. O resumo, sumário e capa só perdem ponto se estiverem ausentes.
ETAPA 1: MAPEAMENTO ESTRUTURAL INICIAL
Leia o relatório inteiro e documente:
Tema do experimento: [identificar]
Seções presentes: [listar todas]
Componentes: Figuras: X | Tabelas: Y | Equações: Z | Referências: W
Primeira Impressão (Neutra): [descrever brevemente o aspecto geral do documento]
ETAPA 2: ANÁLISE DE CONTEÚDO E ESTRUTURA POR SEÇÃO
Para cada seção do relatório, siga este modelo de saída:
AVALIAÇÃO DA SEÇÃO: [NOME DA SEÇÃO]
Critérios Aplicáveis: [listar os critérios da seção]

Nota Máxima Possível: [X,X pontos / ou Penalidade de -Y,Y]
Pontos Positivos: [listar aspectos específicos que estão corretos e bem-feitos]
Análise Crítica e Ações Corretivas:
ERRO #1: [Descrição precisa do erro.] Localização: [Página X...] Ação Corretiva: [Sugestão específica...]
Nota Atribuída / Penalidade Aplicada: [X,Y pontos ou -Z pontos]
Justificativa da Nota: [Explicação concisa da pontuação]
(...O restante do prompt continua, detalhando as etapas de 3 a 8, incluindo os checklists de penalidades técnicas e o método de cálculo final da nota...)"

Com este instrumento (*prompt*) refinado e validado em mãos, iniciamos o processo de avaliação automatizada, confiantes de que a estrutura imposta permitiria uma análise comparativa rigorosa e significativa, cujos critérios quantitativos e qualitativos são detalhados na seção seguinte.

Para medir a divergência entre as notas, utilizamos as seguintes métricas:

Diferença Absoluta (Δ): O valor absoluto da diferença entre a nota humana (N_H) e a nota da IA (N_IA), representada por $\Delta = |N\_H - N\_IA|$.

Erro Relativo: Para contextualizar a magnitude da divergência em relação à avaliação do Avaliador Humano, calculamos o erro relativo, que indica o quão distante a nota da IA está da nota humana em termos percentuais.

Além das notas (números), analisamos a natureza do feedback fornecido por ambas as partes. O foco foi identificar em quais critérios (formatação, rigor científico, clareza conceitual) o Avaliador IA se aproximava ou se distanciava do julgamento do Avaliador Humano. Demos especial atenção às nuances metodológicas, investigando a capacidade de cada um em identificar falhas sutis de raciocínio científico, um domínio onde, hipoteticamente, a avaliação humana se destacaria. Os resultados detalhados desta comparação são apresentados na seção seguinte.

3. Resultados e Discussão

A análise comparativa dos resultados entre as avaliações realizadas pelo Avaliador Humano e pela Inteligência Artificial (IA) foi conduzida com base nas pontuações atribuídas a cada seção de quatro relatórios técnicos científicos, denominados Grupo 1, Grupo 2, Grupo 3 e Grupo 4. As observações e discrepâncias qualitativas no feedback também foram criteriosamente examinadas.

3.1. Avaliação do Grupo 1

A análise das pontuações do Grupo 1, conforme detalhado na Tabela 1, revelou uma notável discrepância inicial entre as avaliações humanas e por IA. Enquanto as seções de Capa, Sumário e Resumo foram pontuadas com 1,0 ponto pela IA (indicando presença e conformidade mínima, visto que eram penalizadas apenas se ausentes), o avaliador humano não atribuiu pontuação explícita, pois sua metodologia considerava essas seções apenas em caso de penalização por ausência.

Tabela 1. Pontuações obtidas pelo Grupo 1

GRUPO 1		
Nota (pontos)		
Seção	Inteligência Artificial (IA)	Avaliador humano
Introdução	0,8	0,1
Fundamentação teórica	0,9	0,2
Procedimento experimental	0,9	0,2
Resultados & Discussão	1,3	0
Conclusão	1,2	0,2
Total de pontos	5,1	0,7
Total de penalidades	-1,0	-
Nota final	4,1	0,7

Entretanto, a Avaliação por IA também incorporava penalidades específicas. A IA aplicou deduções por: citação na Introdução (0,2 pontos), citação na Fundamentação Teórica (0,1 pontos), figuras fora do padrão recorrente (0,2 pontos), valores inconsistentes em tabelas (0,3 pontos) e referências (0,2 pontos), totalizando 1,0 pontos em penalidades.

Em termos de concordância qualitativa, a IA apontou os mesmos erros centrais identificados pelo avaliador humano, como problemas de citação referencial, incoerência teórica e erros gramaticais em outras seções do trabalho. Contudo, a IA ofereceu sugestões adicionais para a Capa (padronização de local e ano) e Sumário (atualização numérica na revisão final), observações que não foram explicitadas nas anotações da Avaliação Humana. Curiosamente, a IA não identificou um erro gramatical ("cronometro" sem acento) pontuado pelo avaliador humano no resumo. A divergência mais marcante no resumo foi que, enquanto a IA não aplicou penalização direta, o avaliador humano considerou o objetivo incoerente com os critérios. Vale realçar que a IA pontuou a Capa, Sumário e Resumo somente do Grupo 1.

Considerando as penalidades, a nota final do Grupo 1, calculada pela IA (Equação 1), foi determinada pela subtração do total de pontos pelas penalidades. A Tabela 1 resume essa comparação, revelando que a nota final da IA foi de 4,1 pontos, enquanto a da Avaliação Humana foi de 0,7 pontos. Essa diferença de 3,4 pontos significa que a nota da Avaliação Humana foi 79,41% menor que a da IA, indicando uma avaliação significativamente mais rigorosa por parte do docente.

3.2. Avaliação do Grupo 2

A análise do Grupo 2 (Tabela 2) seguiu um padrão semelhante de discrepância. A IA desconsiderou as pontuações de Capa, Resumo e Sumário, assim como o avaliador humano, indicando ausência ou não aplicabilidade de penalidades nessas seções. No entanto, houve divergências nos pontos positivos e nas falhas identificadas. Por exemplo, no Resumo, a IA considerou o objetivo como ponto positivo, enquanto o avaliador humano o julgou incoerente. A IA, por sua vez, destacou a ausência de resultados quantitativos, tornando a seção subjetiva.

Tabela 2. Pontuações obtidas pelo Grupo 2

GRUPO 2		
Nota (pontos)		
Seção	Inteligência Artificial (IA)	Avaliador Humano
Introdução	0,5	0,2
Fundamentação teórica	1	0,2
Procedimento experimental	0,5	0,2
Resultados & Discussão	1,5	0,2
Conclusão	0,5	0
Total de pontos	4	0,8
Total de penalidades	-1,2	-
Notas finais	2,8	0,8

Na Introdução, a Avaliação Humana focou na necessidade de apresentar resultados na conclusão e na clareza do objetivo de estimar a aceleração da gravidade. Em contrapartida, a IA identificou uma grave falha conceitual na frase "Então não tem um erro e um acerto para o experimento", salientando que a experimentação física visa definir quantitativamente erros e medir grandezas.

Na Fundamentação Teórica, a Avaliação Humana apontou lacunas e o uso inadequado de pronomes indefinidos. A IA, de forma mais aprofundada, criticou a relevância das teorias citadas (Lei Gravitacional de Newton e Lei de Ação e Reação) para a medição da aceleração da gravidade, sugerindo a inclusão de Teoria de Erros, Tratamento Estatístico de Dados e Cinemática do Movimento Retilíneo Uniforme Variado (MRUV). A IA também identificou uma citação indireta incorreta. No Procedimento Experimental, o avaliador humano absteve-se de comentários, enquanto a IA apontou a ausência de fontes para figuras e a falta de precisão na medição dos instrumentos.

Na seção de Resultados e Discussão, ambos os avaliadores convergiram na identificação de inconsistências nos resultados e numeração errada das equações. A Avaliação Humana notou a ausência de unidades de medida nos valores numéricos, e a IA ressaltou que "a discussão é praticamente inexistente". Por fim, na Conclusão, o avaliador humano atribuiu nota zero sem comentários adicionais. A IA, por sua vez, detalhou que a conclusão não apresentou resultados quantitativos que sustentam as afirmações dos alunos e enfatizou que "uma conclusão científica deve sintetizar os achados numéricos". Conforme a Tabela 2, a nota final da Avaliação Humana (0,8 pontos) foi 71,43% menor que a da IA (2,8 pontos), novamente indicando uma maior severidade do avaliador humano.

3.3. Avaliação do Grupo 3

Para o Grupo 3 (Tabela 3), observou-se uma diferença de 5,55 pontos entre a nota final da IA (9,0 pontos) e a da Avaliação Humana (3,45 pontos). A IA não aplicou penalidades neste relatório, o que contribuiu para sua pontuação elevada. A avaliação humana demonstrou uma interpretação mais abrangente, não se atendo apenas a falhas técnicas, mas também a incoerências conceituais e estruturais. O avaliador humano

apontou inconsistências entre o resumo e o restante do relatório, erros referenciais, ausência de conceitos, lacunas na explicação de equações e aplicação de resultados no procedimento experimental.

Tabela 3. Pontuações obtidas pelo Grupo 3

GRUPO 3		
Nota (pontos)		
Seção	Inteligência Artificial (IA)	Avaliador humano
Introdução	2	1
Fundamentação teórica	2	0,5
Procedimento experimental	0,8	0,2
Resultados & Discussão	2,5	0,75
Conclusão	1,7	1
Total de pontos	9	3,45
Total de penalidades	0	-
Nota final	9	3,45

Apesar de reconhecer que o relatório estava "bem feito", a Avaliação Humana ressaltou a falta de relação com o roteiro proposto. Essa avaliação mais holística e contextualizada resultou em uma nota 61,66% menor que a da IA, que tendeu a elogiar o relatório e a não penalizar os erros com o mesmo rigor, ou com o mesmo tipo de interpretação contextual, demonstrando a influência da subjetividade e da experiência humana em uma correção.

3.4. Avaliação do Grupo 4

No caso do Grupo 4 (Tabela 4), observou-se uma inversão do padrão anterior: a nota da Avaliação Humana (6,75 pontos) foi maior que a da IA (5,3 pontos). O avaliador humano fez apenas dois comentários pontuais: um na introdução, sugerindo "situar o leitor a respeito do tema" e "deixar claro o principal objetivo da atividade", e outro em relação ao uso inadequado do termo "rápido" para tempo. Nas demais seções, o avaliador humano absteve-se de comentários.

A IA considerou o relatório "interessante" e forneceu feedback detalhado. Na Introdução, apontou que, apesar do bom objetivo, a descrição confusa e erros no sumário indicavam falta de revisão. A Fundamentação Teórica foi elogiada como uma das melhores análises. No Procedimento Experimental, destacou o detalhamento e declaração de incertezas instrumentais, penalizando apenas a ausência de fonte nas figuras (0,5 pontos). Nos Resultados e Discussão, reconheceu o desenvolvimento conceitual, mas criticou a falha em conectar com a realidade física. A Conclusão foi considerada "muito fraca e desconectada", não abordando o valor "absurdo" de 'g' encontrado. Este foi considerado o trabalho mais completo comparado aos anteriores. A nota humana foi 21,4% maior que a da IA sugerindo que em relatórios de maior qualidade, a IA pode ser mais crítica que o avaliador humano.

Tabela 4. Pontuações obtidas pelo Grupo 4

GRUPO 4		
Nota (pontos)		
Seção	Inteligência Artificial (IA)	Avaliador humano
Introdução	1,5	0,5
Fundamentação teórica	2	1,5
Procedimento experimental	0,8	0,75
Resultados & Discussão	1	2,5
Conclusão	0,5	1,5
Total de pontos	5,8	6,75
Total de penalidades	-0,5	-
Nota final	5,3	6,75

A análise dos quatro relatórios demonstra que a IA, embora capaz de corrigir baseada em critérios detalhados, não alcança a profundidade de avaliação humana. As discrepâncias nas notas e *feedback* qualitativo indicam que a IA se destaca na identificação de erros formais e aplicação consistente de regras, mas a interpretação de conceitos complexos, identificação de lacunas contextuais e compreensão da intenção pedagógica permanecem domínios onde a avaliação humana mantém superioridade.

4. Considerações finais

Os resultados deste relato reforçam que a Inteligência Artificial, embora valiosa no apoio à avaliação de relatórios acadêmicos, não deve ser utilizada como única fonte de correção. As divergências identificadas evidenciam limites da abordagem automatizada. A análise dos quatro grupos revelou um padrão consistente: a IA apresentou forte desempenho na aplicação de regras formais e detecção de erros estruturais, enquanto o avaliador humano destacou-se na interpretação de sutilezas metodológicas, análise da coerência conceitual e profundidade argumentativa.

Essa diferença justifica as discrepâncias nas notas, como a diferença de 5,55 pontos no Grupo 3. Mesmo diante dos mesmos problemas, IA e avaliador atribuíam pesos distintos aos critérios. Conclui-se que a IA representa um recurso promissor para automatizar tarefas avaliativas objetivas e fornecer feedback padronizado. Contudo, o olhar crítico e pedagógico do docente permanece essencial para garantir uma avaliação formativa completa e sensível às particularidades do aprendizado.

Investigações futuras devem priorizar múltiplas frentes de desenvolvimento. Primeiramente, a análise da percepção docente sobre a viabilidade da IA como ferramenta de apoio à correção, seguida da exploração de diferentes configurações paramétricas para otimização do processo. Estudos com amostras maiores e abordagens híbridas que integrem eficiência tecnológica e supervisão pedagógica humana também representam direções essenciais, incluindo a validação da metodologia em disciplinas correlatas como Química e Biologia para estabelecer a generalização da abordagem além da Física Experimental. Adicionalmente, a investigação de *fine-tuning* de modelos de linguagem para contextos educacionais específicos e a inclusão da perspectiva discente sobre o uso de IA na avaliação acadêmica constituem linhas de pesquisa fundamentais para uma compreensão mais abrangente do impacto desta tecnologia no processo educativo.

Agradecimentos

Agradeço ao Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP) - Campus São José dos Campos pela concessão da bolsa no Programa Institucional de Bolsas de Iniciação Científica e Tecnológica do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo. Iniciação em Desenvolvimento Tecnológico e Inovação (PIBITI) que viabilizou esta pesquisa e pelo suporte institucional.
Declaro o uso de IA Generativa, Gemini AI Studio, para auxiliar na estrutura inicial, revisão gramatical e para analisar se as seções dos relatórios estão de acordo com os critérios estabelecidos pela organização do evento.

4. Referência

ARAÚJO, M. S. T., & ABIB, M. L. V. S. (2003). Atividades experimentais no ensino de física: diferentes enfoques, diferentes finalidades. *Revista Brasileira de Ensino de Física,* 25(2), 176-194. https://www.scielo.br/j/rbef/a/PLkjm3N5KjnXKgDsXw5Dy4R/?lang=pt. Junho.

BARRETO, R., & COSTA, T. S. (2017). A avaliação da aprendizagem no ensino superior: limites e possibilidades. *Revista Práxis Educacional*, 13(26), 109–130. https://periodicos2.uesb.br/index.php/praxis.

BLACK, P., & WILIAM, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7-74. https://doi.org/10.1080/0969595980050102. Março.

BROWN, S., & KNIGHT, P. (2012). Assessing learners in higher education. Routledge. https://eric.ed.gov/?id=ED369379.

LIMA, R. M., et al. (2017). Desenvolvimento de competências em engenharia: uma abordagem baseada em projetos. *Educação em Revista*, 33, e172508. https://www.scielo.br/j/edur/.

MOK, R., et al. (2024). Using AI Large Language Models for Grading in Education: A Hands-On Test for Physics. *arXiv*. https://arxiv.org/abs/2411.13685.

NUNES, A. A., et al. (2021). Rubricas de avaliação como instrumento pedagógico: reflexões sobre equidade e aprendizagem significativa. *Revista de Educação da Universidade Federal do Vale do São Francisco*, 11(23), 153–171. http://www.periodicos.univasf.edu.br/.

REYNOLDS, L., & MCDONELL, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-7). https://arxiv.org/abs/2102.07350.

WANG, X., et al. (2023). Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/pdf?id=1PL1NIMMrw.

WEI, J., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837. https://arxiv.org/abs/2201.11903.