Inteligência Artificial Aplicada à Análise das Habilidades Centrais nas Provas Teóricas da OBR (2014–2023)

Diogo de Oliveira Lima
Faculdade de Computação
Universidade Federal do MS
Campo Grande, Brasil
diogo_lima@ufms.br

Gabriel Alves Massuda
Faculdade de Computação
Universidade Federal do MS
Campo Grande, Brasil
gabriel.massuda@ufms.br

Amaury Antonio de Castro Junior
Faculdade de Computação
Universidade Federal do MS
Campo Grande, Brasil
amaury.junior@ufms.br

***Abstract*—This paper presents a pedagogical analysis of the theoretical exams of the Brazilian Robotics Olympiad (OBR) from 2014 to 2023, focusing on the core skills required according to the Base Nacional Comum Curricular (BNCC). Using the GPT-4o model, questions were classified by subject area, curriculum component, and competition level. The results highlight recurring patterns and support the development of more targeted educational materials.**

***Index Terms*—OBR, lógica, GPT-4o, análise de prova, BNCC, educação básica, olimpíadas científicas.**

I. Introdução

A Olimpíada Brasileira de Robótica (OBR) tem se consolidado como uma das principais iniciativas nacionais de estímulo à ciência, tecnologia e raciocínio lógico entre estudantes da educação básica. Sua modalidade teórica propõe desafios interdisciplinares que exigem interpretação, lógica e conhecimentos alinhados à Base Nacional Comum Curricular (BNCC), tornando-se um ambiente fértil para o desenvolvimento cognitivo e o engajamento escolar.

Neste contexto, este artigo investiga como as competências avaliadas pela OBR, ao longo de suas edições teóricas entre **2014 e 2023**, dialogam com as habilidades previstas na BNCC. A proposta envolve uma análise automatizada e estruturada das provas, com o objetivo de identificar padrões pedagógicos recorrentes e contribuir para o aprimoramento de práticas educacionais voltadas à preparação de estudantes para a competição.

II. Fundamentos sobre Inteligência Artificial

A metodologia deste estudo emprega recursos de Inteligência Artificial (IA) baseados em modelos de linguagem. Para garantir clareza a leitores não especializados em IA, esta seção apresenta uma explicação introdutória sobre os principais conceitos e termos técnicos adotados.

A. Modelos de linguagem e o GPT

Modelos de linguagem são sistemas de IA treinados para compreender e gerar texto em linguagem natural. Eles identificam padrões estatísticos em grandes volumes de dados textuais, sendo capazes de resumir informações, classificar trechos e responder perguntas.

O termo **GPT** significa *Generative Pre-trained Transformer*. Trata-se de uma família de modelos desenvolvidos pela OpenAI com base na arquitetura *Transformer*, que permite analisar sequências de palavras considerando dependências de longo alcance. O modelo utilizado neste estudo foi o **GPT-4o**, lançado em 2024, que representa uma versão otimizada e multimodal da quarta geração do GPT, capaz de interpretar tanto texto quanto imagens. O termo **snapshot** indica a versão estável do modelo disponível em uma data específica (neste caso, maio de 2025).

B. Parâmetros de inferência

Na utilização de modelos generativos, alguns parâmetros controlam o comportamento das respostas:

- **Temperatura = 0,2**: regula o grau de variabilidade ou aleatoriedade da resposta. Valores próximos de 0 tornam as saídas mais determinísticas e consistentes; valores maiores aumentam a diversidade. Neste estudo, o valor baixo (0,2) foi escolhido para priorizar consistência nas classificações.
- **Top-p = 1,0**: define o limite de probabilidade cumulativa das palavras candidatas. O valor 1,0 equivale a não restringir o vocabulário, sendo usado aqui para preservar a completude das respostas.
- **Max_tokens = 800**: estabelece o número máximo de unidades de texto (tokens) que o modelo pode gerar em uma resposta. Cada token corresponde aproximadamente a uma palavra ou parte de palavra. O limite de 800 tokens foi adotado para assegurar respostas completas sem ultrapassar a capacidade do modelo.

C. Ferramentas de processamento de texto

As provas da OBR foram disponibilizadas em arquivos PDF. Para extrair seu conteúdo em formato textual, empregou-se a biblioteca `pdfplumber`, que converte páginas em texto limpo, preservando a estrutura básica de parágrafos e tabelas.

Posteriormente, expressões regulares (`regex`) foram aplicadas para padronizar trechos e remover inconsistências de formatação.

D. Estratégia de classificação

O processo de classificação automática neste estudo se baseou no uso de instruções fornecidas ao modelo, conhecidas como **prompts**. Em termos gerais, prompts funcionam como comandos que orientam a IA sobre qual tarefa deve ser realizada. Também foi utilizada a técnica de **few-shot learning**, em que exemplos de entrada e saída são incluídos no prompt para guiar o modelo e reduzir respostas genéricas.

A consistência das classificações foi avaliada por meio de duas métricas principais: **concordância percentual** e **coeficiente de Cohen** (κ). A concordância percentual mede simplesmente a proporção de vezes em que as classificações realizadas pela IA coincidiram com as feitas por avaliadores humanos, sendo, portanto, uma medida intuitiva e de fácil compreensão. O coeficiente de Cohen, por sua vez, quantifica o grau de concordância *além do acaso*. Essa estatística corrige a possibilidade de coincidências aleatórias: valores próximos de 0 indicam baixa consistência, entre 0,40 e 0,60 sugerem concordância moderada, e acima de 0,80 são interpretados como concordância quase perfeita. O uso dessas métricas em conjunto é amplamente reconhecido em pesquisas educacionais e nas ciências sociais, pois permite avaliar não apenas a taxa bruta de coincidências, mas também a robustez da concordância entre avaliadores independentes.

E. Visualizações

A consolidação dos resultados foi realizada no Google Sheets, mas algumas figuras (como o mapa de calor) foram geradas com auxílio de um *script* Python utilizando a biblioteca `matplotlib`, ferramenta amplamente empregada na comunidade científica para criação de gráficos bidimensionais.

F. Escopo da análise

Cabe destacar que o foco desta pesquisa concentrou-se em análises **quantitativas**, como contagens, distribuições e gráficos. Não foram produzidos *textos analíticos* individuais para cada prova ou questão, pois o objetivo era mapear padrões estatísticos em larga escala, e não discutir qualitativamente cada enunciado.

G. Relevância da explicitação dos termos

A explicitação dos parâmetros, ferramentas e estratégias acima descritas contribui para a transparência e a reprodutibilidade do estudo. Dessa forma, mesmo leitores sem formação específica em Inteligência Artificial podem compreender as escolhas metodológicas e situar melhor o papel do GPT-4o e das técnicas complementares no processo de análise automatizada.

III. Metodologia

A abordagem metodológica foi organizada em quatro etapas: (i) coleta e pré-processamento dos enunciados; (ii) classificação automática via GPT-4o com ajuste iterativo de *prompt*; (iii) consolidação estatística e visual; e (iv) proposição de validação humana como trabalho futuro. Todo o fluxo segue recomendações de transparência em IA educacional [6] e de mitigação de viés [7], [9].

A. Coleta e pré-processamento

As provas teóricas da OBR (2014–2023, níveis 0 – 5) foram baixadas em PDF do repositório oficial[1] e anexadas à sessão de conversa com o *ChatGPT*, permitindo acesso multimodal interno. Utilizando `pdfplumber` e expressões regulares, cada página foi convertida em texto limpo e numerada. Obtiveram-se **835** questões válidas, distribuídas em 342 de Ciências da Natureza, 290 de Matemática, 126 de Linguagens e 77 de Ciências Humanas. Os dados foram armazenados em planilhas Google em duas camadas:

1) **Tabelas anuais niveladas**: 60 planilhas (10 anos × 6 níveis) com colunas *ID, enunciado, gabarito, área, componente, habilidade*;
2) **Tabela global**: soma de todas as questões por área, base para a Fig. 1.

B. Classificação automática e ajuste de prompt

Empregou-se o modelo `gpt-4o` (OpenAI, snapshot 13 mai 2025) com temperatura = 0,2, *top-p* = 1,0 e *max_tokens* = 800. Cada questão recebeu o **Prompt 1**:

> Categorize a questão em: (i) Área BNCC; (ii) Componente Curricular; (iii) Habilidade central exigida. Analise cuidadosamente o enunciado; evite respostas genéricas.

Na rodada-piloto (provas 2014–2015) a concordância IA × humano foi 68 % ($\kappa \approx 0{,}54$). Três iterações elevaram-na a 92 % ($\kappa \approx 0{,}86$) numa amostra cega de 84 questões (10 % estratificado):

1) inclusão de exemplos *few-shot*;
2) regras condicionais (p. ex. "Se o enunciado contiver *força* ou *aceleração*, classificar em Física");
3) filtro regex que rejeitava termos vagos e reenviava o item em temperatura = 0,0.

Após atingir esse limiar, as provas 2016–2023 foram processadas sem intervenção adicional.

Observação: não foram produzidos textos analíticos por ano; o foco concentrou-se na classificação quantitativa e na geração de tabelas e gráficos.

C. Consolidação estatística e visualização

A consolidação final ocorreu inteiramente no **Google Sheets**. As planilhas anuais foram combinadas em uma aba síntese (totais por área e por nível). A partir dessa aba geraram-se:

- **Tabela I** – totais 2014–2023, criada com a opção "Inserir → Tabela" do Sheets;

[1] https://obr.robocup.org.br/documentos-e-manuais/

- **Gráficos de pizza acumulativos por nível** (Figs. 3–8) – ferramenta nativa de gráficos, tipo "Pizza";
- **Mapa de calor nível × área** (Fig. 2) – gerado via ChatGPT, que executou um *script* Python utilizando `matplotlib` e exportou o resultado em PDF;

Essa combinação de Google Sheets (para tabelas e gráficos) e ChatGPT + `matplotlib` (para o mapa de calor) permitiu produzir todas as visualizações sem recorrer a bibliotecas locais como `Pandas` ou `NumPy`.

D. Validação humana (trabalho futuro)

Até o presente, não houve revisão especializada externa. Planeja-se, em estudos subsequentes, convidar docentes de cada área da BNCC para avaliar cerca de 10 % das questões de forma cega e independente, calculando concordância percentual e coeficiente de Cohen [8] (meta $\kappa \geq 0{,}80$). Divergências recorrentes servirão de exemplos *few-shot* e regras condicionais nos próximos ciclos de ajuste.

IV. Resultados

Os resultados consolidados das provas teóricas da OBR (2014–2023) revelam padrões consistentes de distribuição entre as áreas da BNCC. As Tabelas I–II e as Figuras 1–2 mostram que Ciências da Natureza (41,0%) e Matemática (34,7%) concentram cerca de três quartos das 835 questões analisadas, enquanto Linguagens (15,1%) e Ciências Humanas (9,2%) aparecem em proporções reduzidas. Esse desbalanceamento evidencia a centralidade das áreas STEM na prova, mas também aponta uma sub-representação sistemática das competências ligadas à leitura crítica, à comunicação e à contextualização sociocultural — dimensões igualmente previstas na BNCC.

Tabela I
Quantidade total de questões de cada área (2014-2023)

Área da BNCC	total de questões
Ciências da Natureza	**342**
Matemática	**290**
Linguagens	126
Ciências Humanas	77

Tabela II
Total de questões em cada nível (soma 2014–2023)

Nível	Natureza	Matemática	Linguagens	Humanas
0	31	47	16	6
1	32	43	19	6
2	52	60	17	16
3	54	54	27	15
4	58	52	24	16
5	115	34	23	18

A análise acumulada por nível, apresentada nas Figuras 3–8, aprofunda esse cenário. Nos níveis iniciais (0–2), a predominância da Matemática revela o papel estruturante do raciocínio lógico e algébrico como porta de entrada da

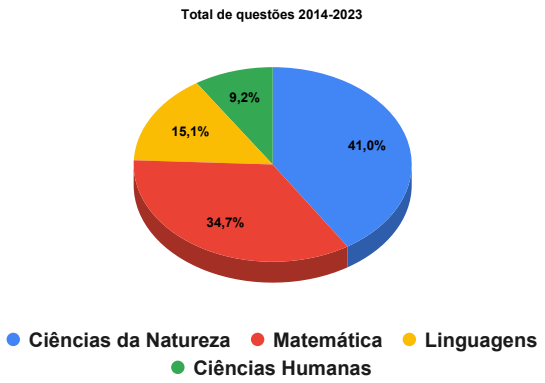

Fig. 1. Distribuição global de questões por área da BNCC (soma 2014–2023, todos os níveis).

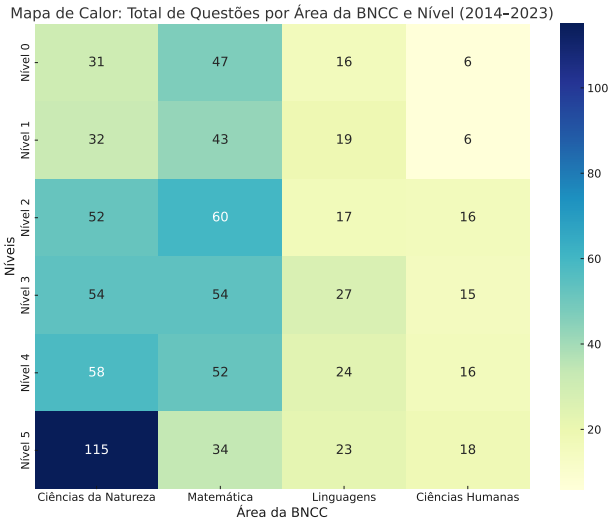

Fig. 2. Mapa de calor — *total* de questões por área da BNCC e por nível (soma 2014–2023).

competição. A partir do Nível 3, observa-se uma virada: Ciências da Natureza alcança equilíbrio com a Matemática e passa a ganhar espaço progressivamente. Nos níveis mais avançados (4–5), a ênfase desloca-se de forma clara para conteúdos de Física e Biologia, enquanto Matemática perde protagonismo e Linguagens e Humanas permanecem secundárias. Esse movimento sugere que a OBR estrutura suas avaliações como uma progressão pedagógica: inicia-se pela lógica matemática elementar, transita para o equilíbrio entre áreas e culmina em competências científicas aplicadas, alinhadas às habilidades da BNCC do Ensino Fundamental II e Ensino Médio.

Além disso, chama atenção o baixo crescimento proporcional de Linguagens e Humanas, que mesmo nos níveis superiores não ultrapassam um quarto do total de itens. Essa permanência em patamares modestos pode reforçar um currículo oculto em que a argumentação textual e a análise crítica de transformações sociais são menos valorizadas na

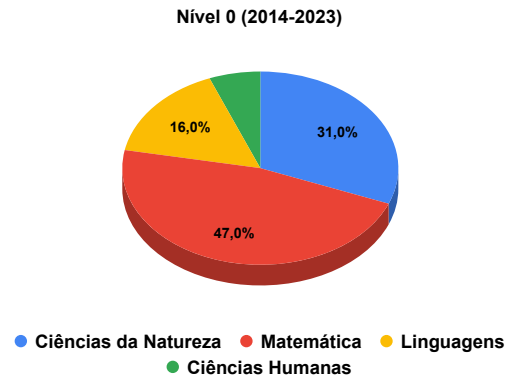

Fig. 3. Nível 0 — totais acumulados de questões por área da BNCC (2014–2023).

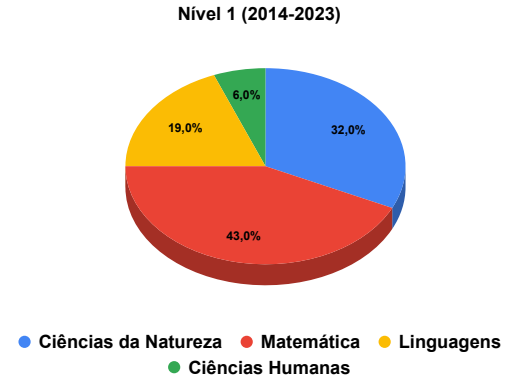

Fig. 4. Nível 1 — totais acumulados de questões por área da BNCC (2014–2023).

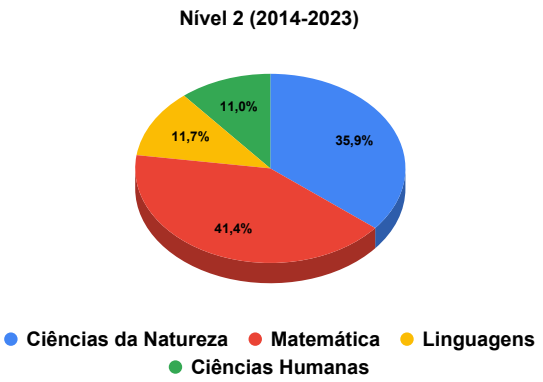

Fig. 5. Nível 2 — totais acumulados de questões por área da BNCC (2014–2023).

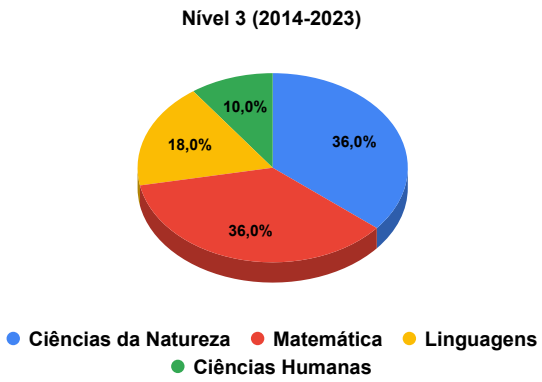

Fig. 6. Nível 3 — totais acumulados de questões por área da BNCC (2014–2023).

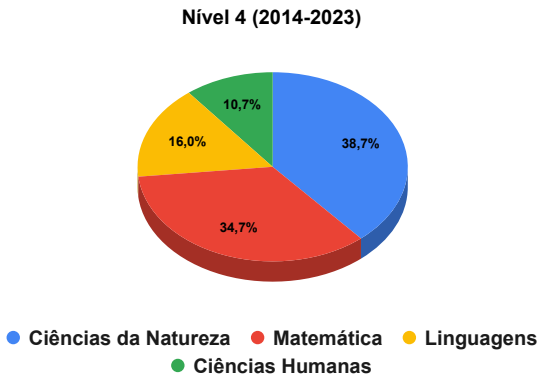

Fig. 7. Nível 4 — totais acumulados de questões por área da BNCC (2014–2023).

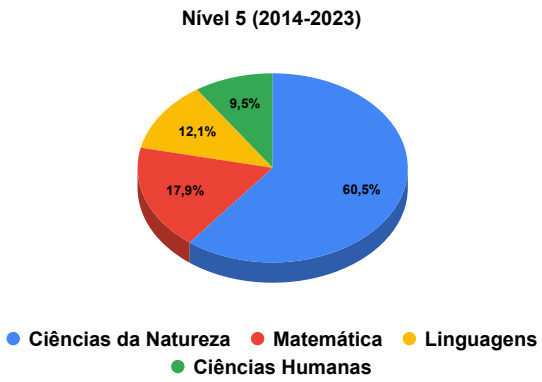

Fig. 8. Nível 5 — totais acumulados de questões por área da BNCC (2014–2023).

competição. Ainda que a OBR seja reconhecida como evento de estímulo científico, a ausência de maior interdisciplinaridade pode limitar a contribuição da olimpíada para uma formação integral, conforme orienta a BNCC.

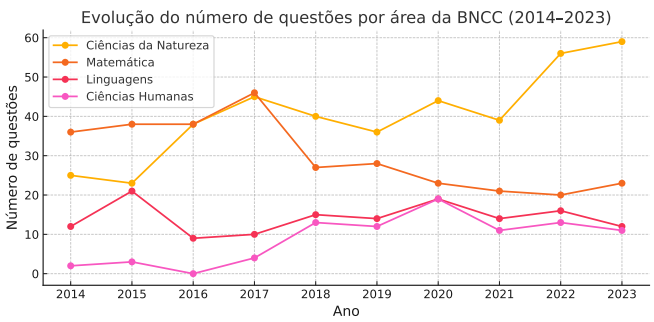

Fig. 9. Evolução anual do número de questões por área da BNCC (níveis 0–5 agregados).

A tendência temporal consolidada (Fig. 9) reforça essa interpretação. Entre 2014 e 2023, Ciências da Natureza praticamente dobra sua participação anual, passando de 25 para 59 itens, enquanto Matemática recua de 36 para 23. Linguagens e Humanas apresentam relativa estabilidade, mas com pico em 2020. O crescimento contínuo da área científica sugere também uma intencionalidade do comitê de prova em valorizar temas aplicados à robótica, como energia, sensores e fenômenos físicos, o que fortalece o vínculo com STEM, mas enfraquece a diversidade curricular prevista pela BNCC.

Em síntese, os dados indicam três movimentos centrais: (i) a Matemática domina os níveis iniciais, mas perde espaço gradualmente; (ii) Ciências da Natureza assume protagonismo nos níveis avançados e cresce de forma consistente ao longo dos anos; e (iii) Linguagens e Humanas permanecem secundárias, com participação pontualmente ampliada apenas em condições específicas. Essa configuração revela não apenas o perfil das provas, mas também possíveis impactos na preparação de estudantes e no planejamento pedagógico de escolas que utilizam a OBR como referência.

V. Discussão

Esta seção interpreta criticamente os achados quantitativos, discute implicações pedagógicas de longo prazo e explicita limitações metodológicas e riscos de viés.

A. Assimetria entre áreas da BNCC

Evolução curricular ao longo da década.: A Figura 9 revela uma inflexão importante no foco da OBR: Ciências da Natureza salta de 25 itens em 2014 para 59 em 2023 (+34), enquanto Matemática recua de 36 para 23 (–13). A curva de Linguagens e Humanas exibe pico conjunto em 2020 — edição remota em que questões de interpretação e contexto sociotécnico ganharam espaço — mas volta a patamares modestos nos três anos seguintes. Esses deslocamentos sugerem que o comitê de prova vem privilegiando abordagens experimentais ligadas a sensores, eletrônica e biotecnologia, potencialmente à custa de competências de modelagem matemática e argumentação histórico-cultural.

A Tabela I mostra que **Ciências da Natureza (41,0 %)** e **Matemática (34,7 %)** concentram três quartos das 835 questões analisadas, ao passo que Linguagens (15,1 %) e Ciências Humanas (9,2 %) somam apenas um quarto. Essa desigualdade persiste em todos os níveis (Fig. 2) e atinge o ápice no Nível 5, onde Natureza representa $\frac{115}{190} \approx 60\,\%$ dos itens (Fig. 8).

Para mensurar o desbalanceamento, calculou-se o índice de diversidade de Simpson: $D = 1 - \sum p_i^2 = 0{,}68$,[2] valor que confirma concentração moderada em torno de duas grandes áreas.

Possíveis causas: (i) natureza técnico-científica da robótica, que demanda forte base matemática e física; (ii) composição do comitê de prova, tradicionalmente formado por docentes de Engenharia e Ciências Exatas; (iii) escassez de questões interdisciplinares já prontas em bancos públicos.

B. Impactos pedagógicos de longo prazo

- **Currículo oculto.** A escassez de itens de Linguagens e Humanas sinaliza, ainda que implicitamente, que competências de leitura crítica, comunicação científica e reflexão ética são menos relevantes para o sucesso na prova.
- **Preparação assimétrica.** Escolas que usam a OBR como base de estudo tendem a priorizar conteúdos de Física, Eletricidade e Álgebra, em detrimento de práticas discursivas, cidadania digital ou análise sociocultural — também previstas na BNCC. A habilidade `EF09LP14`, por exemplo, envolve a produção e análise de textos argumentativos, com foco em justificar teses e avaliar a estrutura de argumentos. Já a `EM13CHS502` propõe a análise crítica das transformações no mundo do trabalho e seus impactos sociais. Tais competências, essenciais à formação integral, são pouco abordadas nos itens da OBR analisados.
- **Avaliação formativa limitada.** A ausência de itens que articulem robótica a contextos históricos ou sociais dificulta diagnosticar o desenvolvimento integral dos estudantes, restringindo-o ao eixo STEM tradicional.

Para mitigar esses efeitos, recomenda-se que futuras provas incluam questões que abordem, por exemplo, o impacto da automação no trabalho, a acessibilidade tecnológica ou dilemas éticos relacionados à inteligência artificial. Esses temas permitiriam ampliar a distribuição entre as áreas da BNCC e estimular competências críticas e interdisciplinares, alinhadas à formação integral proposta pelo currículo.

C. Transparência e mitigação de viés

A metodologia empregada foi guiada por princípios de transparência no uso de IA em contextos educacionais, conforme as diretrizes propostas por Zawacki-Richter et al. [6].

[2]Onde p_i é a fração de questões da área i; $D = 0$ indica domínio absoluto de uma área e $D \to 1$ distribuição perfeitamente uniforme.

O modelo utilizado (`gpt-4o`, maio/2025) teve todos os parâmetros explicitados (temperatura = 0,2, *top-p* = 1,0, *max_tokens* = 800), e o *prompt* aplicado é reproduzido integralmente neste artigo, assegurando reprodutibilidade e rastreabilidade.

Para mitigar vieses, adotaram-se práticas alinhadas a Binns [7], como filtros de expressões regulares, ajustes de temperatura e inserção de exemplos *few-shot*. Tais medidas visam reduzir classificações genéricas e promover consistência. A validação cega por especialistas — proposta como trabalho futuro — complementará esse esforço, permitindo refinar o modelo com base em julgamentos humanos.

Embora Bybee [9] trate da equidade em STEM de forma mais ampla, seus princípios reforçam a crítica central deste trabalho: a necessidade de maior equilíbrio entre áreas do conhecimento, incluindo temas socioculturais e éticos nas provas da OBR.

D. Limitações do estudo

Além das limitações já mencionadas na seção anterior, destacam-se os seguintes pontos que devem ser considerados na interpretação dos resultados:

- **Extração de texto:** O processo de conversão dos arquivos PDF para texto limpo pode ter gerado perdas de conteúdo, especialmente em elementos gráficos, tabelas e diagramas que, porventura, compunham o corpo das questões originais.
- **Classificação centrada na BNCC:** O estudo optou por adotar exclusivamente a taxonomia da BNCC como critério de categorização das questões. Embora esse recorte ofereça uma referência normativa importante, outras matrizes cognitivas, como a Taxonomia de Bloom ou frameworks de avaliação internacional como o PISA, poderiam oferecer análises complementares sobre os níveis de complexidade cognitiva envolvidos.
- **Ausência de validação humana externa:** Apesar da alta concordância interna ($\kappa \approx 0,86$) alcançada nas amostras de teste, a ausência de uma validação cega por especialistas de diferentes áreas limita a generalização dos resultados.
- **Foco exclusivo na prova teórica:** A pesquisa contemplou apenas a modalidade teórica da OBR, desconsiderando as dimensões procedimentais, atitudinais e práticas envolvidas na competição.
- **Análise exclusivamente quantitativa:** A abordagem restringiu-se a contagens e distribuições, sem explorar qualitativamente a natureza cognitiva ou pedagógica das questões, o que pode reduzir a profundidade interpretativa dos achados.

VI. Considerações Finais

Este estudo aplicou inteligência artificial generativa (GPT-4o) para analisar quantitativamente as provas teóricas da Olimpíada Brasileira de Robótica (OBR) entre 2014 e 2023. O mapeamento evidenciou concentração de questões em Ciências da Natureza e Matemática, enquanto Linguagens e Ciências Humanas permanecem minoritárias em todos os níveis. Esse perfil sugere uma trajetória que parte do raciocínio lógico-matemático nos níveis iniciais e culmina em competências científicas aplicadas nos níveis mais avançados.

Esses resultados trazem implicações pedagógicas relevantes. A configuração atual reforça um "currículo oculto" que privilegia conteúdos técnico-científicos em detrimento de competências argumentativas, éticas e socioculturais previstas na BNCC. Para as escolas que utilizam a OBR como referência curricular, tal assimetria pode direcionar a preparação dos estudantes de forma restrita ao eixo STEM.

Como perspectiva futura, recomenda-se ampliar a diversidade temática das provas, incorporando questões que articulem ciência, tecnologia e contexto social de maneira mais equilibrada. Estudos posteriores poderão explorar a modalidade prática da competição, realizar validações independentes com especialistas e investigar impactos de longo prazo na formação dos participantes. Em síntese, o trabalho destaca o potencial da IA para apoiar análises educacionais em larga escala e abre caminho para avaliações mais alinhadas aos princípios de integralidade e equidade curricular.

Acknowledgment

Os autores agradecem à Universidade Federal do Mato Grosso do Sul (UFMS) pelo apoio institucional e pela infraestrutura oferecida para o desenvolvimento desta pesquisa.

References

[1] Olimpíada Brasileira de Robótica (OBR). Provas Anteriores. Disponível em: https://obr.robocup.org.br/documentos-e-manuais/. Acesso em: jun. 2025.

[2] Brasil. Ministério da Educação. Base Nacional Comum Curricular (BNCC). Brasília: MEC, 2018. Disponível em: http://basenacionalcomum.mec.gov.br/. Acesso em: jun. 2025.

[3] MARTINS, Wellington Santos. *Jogos de Lógica: divirta-se e prepare-se para a OBI*. São Paulo: Editora Livros Ilimitados, 2018.

[4] OpenAI. GPT-4o Technical Report, 2024. Disponível em: https://openai.com/index/hello-gpt-4o/. Acesso em: jun. 2025.

[5] ALMEIDA, L. A.; SILVA, R. T. Inteligência artificial na educação básica: desafios e possibilidades. *Revista Brasileira de Informática na Educação*, v. 29, n. 1, p. 245–264, 2023.

[6] ZAWACKI-RICHTER, O.; MARÍN, V. I.; BONN, S.; TABELA, G. Artificial intelligence in higher education: A systematic literature review. *International Journal of Educational Technology in Higher Education*, v. 16, n. 1, p. 1–27, 2019.

[7] BINNS, R. Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, p. 149–159, 2018.

[8] COHEN, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, v. 20, n. 1, p. 37–46, 1960.

[9] BYBEE, R. W. *The Case for STEM Education: Challenges and Opportunities*. NSTA Press, 2013.