Classificação de Questões da Olimpíada Brasileira de Informática Modalidade Iniciação com Inteligência Artificial e Chain of Thought

Thiago Gonçalves de Almeida[1], **Esteic Janaina Santos Batista**[1], **Anderson Correa de Lima**[1], **Amaury Antônio Castro Junior**[1]

[1]Faculdade de Computação - Universidade Federal de Mato Grosso do Sul (UFMS)
Caixa Postal 549, 79.070-900, Campo Grande – MS, Brasil

{almeida_thiago, esteic.batista, anderson.lima, amaury.junior}@ufms.br

Abstract. *This work presents a study on the automatic classification of questions from the Iniciação (Initiation) modality of the Brazilian Olympiad in Informatics (OBI), aiming to compare supervised and prompting-based approaches in order to support the production of training materials for the competition and the development of skills related to Computational Thinking. The methodology was structured into three main stages: (i) mining and extracting text from official OBI exams, resulting in a dataset composed of textual questions; (ii) supervised training using the BERTimbau model; and (iii) prompting-based classification using Chain of Thought (CoT) technique applied to the GPT-4.1-mini and GPT-5-mini models. The results indicate that prompting with step-by-step reasoning is a promising alternative for educational text-classification tasks, reducing the dependency on manual labeling. As a contribution, the study develops an OBI question classifier and provides an associated dataset. This proposal provides a structured repository of educational materials that supports student training and enhances teaching practice, contributing to initiatives for Computer Science education in Basic Education.*

Resumo. *Este trabalho apresenta um estudo sobre a classificação automática de questões da Modalidade Iniciação da Olimpíada Brasileira de Informática (OBI), com o objetivo de comparar abordagens supervisionadas e baseadas em prompting, visando apoiar na produção de materiais de treinamento para a competição e o desenvolvimento de habilidades relacionadas ao Pensamento Computacional. A metodologia foi estruturada em três etapas principais: (i) mineração e extração de textos de provas oficiais da OBI, resultando em um dataset de questões textuais; (ii) treinamento supervisionado com o modelo BERTimbau; e (iii) classificação com prompting por meio da técnica Chain of Thought (CoT), aplicada aos modelos GPT-4.1-mini e GPT-5-mini. Os resultados indicam que o uso de prompting com raciocínio passo a passo é uma alternativa promissora para tarefas educacionais de classificação textual, reduzindo a dependência de rotulação manual. Como contribuição, o estudo desenvolve um classificador de questões da OBI e disponibiliza um dataset associado. Essa proposta oferece um acervo estruturado de material que atende ao treinamento de estudantes e subsidia a prática docente, contribuindo para iniciativas de ensino de Computação na Educação Básica.*

1. Introdução

O conceito de Pensamento Computacional (PC) como habilidade fundamental para todos, foi proposto por [Wing 2016], segundo a autora, o PC compreende um conjunto de atitudes e habilidades tão essenciais quanto a leitura, a escrita e a matemática, devendo ser incorporado como uma competência analítica indispensável na formação de crianças e jovens. Iniciativas de inserção do PC nos currículos tem ganhado destaque em nível global, impulsionada pela expansão das tecnologias digitais e por sua aplicação transversal em diversas áreas do conhecimento [Brackmann et al. 2019].

No cenário internacional, destaca-se o desafio *Bebras*, uma competição educacional voltada ao desenvolvimento do Pensamento Computacional entre estudantes do ensino fundamental e médio [Dagiene and Stupuriene 2016]. Suas atividades abordam conceitos da computação por meio da resolução de tarefas em computador que envolvem representar dados por abstrações, aplicar pensamento algorítmico para automatizar processos, avaliar soluções e generalizar procedimentos para outros problemas.

Em âmbito nacional, a Base Nacional Comum Curricular (BNCC) incorporou a Computação como área de conhecimento, definindo competências que devem ser desenvolvidas ao longo da Educação Básica [Brasil 2022]. Em complemento, a Lei nº 14.533/2023, que institui a Política Nacional de Educação Digital (PNED), reforça a importância do desenvolvimento dessas competências e estabelece diretrizes para formação docente, adequação estrutural das escolas e produção de recursos educacionais voltados à área [Brasil 2023].

Apesar dos avanços normativos, ainda há carência de materiais que apoiem o ensino do Pensamento Computacional na Educação Básica. Dessa forma, as questões da Modalidade Iniciação da Olimpíada Brasileira de Informática (OBI) podem ser uma alternativa interessante para esse propósito, pois apresentam desafios de lógica e raciocínio computacional sem exigir conhecimentos prévios de programação [Instituto de Computação - Unicamp 2025].

Nessa perspectiva, uma Revisão Sistemática da Literatura (RSL) conduzida por [de Almeida et al. 2024], ao investigar sobre estratégias, conceitos de computação e habilidades desenvolvidas em treinamentos de participantes para a competição, identificou que a maioria dos recursos está voltada à Modalidade Programação, enquanto há escassez de materiais direcionados à Modalidade Iniciação. O estudo destacou a necessidade de produzir recursos alinhados às diretrizes da BNCC Computação.

Entre os materiais disponíveis, destaca-se o livro *Jogos de Lógica* [Martins 2011], que apresenta técnicas de resolução de problemas aplicáveis às questões da OBI, envolvendo raciocínio lógico, análise de cenários com regras condicionais, além de ordenação e agrupamento de objetos. Essas habilidades estão diretamente relacionadas ao desenvolvimento do Pensamento Computacional, conforme previsto na BNCC [Brasil 2022].

A partir das lacunas identificadas na RSL citada, este trabalho apresenta o desenvolvimento de um classificador automático de questões da Modalidade Iniciação da OBI, utilizando a técnica de *Chain of Thought* aplicada a grandes modelos de linguagem (*Large Language Models – LLMs*). Como contribuição, o estudo disponibiliza materiais que podem apoiar tanto o treinamento para a competição quanto iniciativas de ensino de Computação na Educação Básica.

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta os fundamentos teóricos e trabalhos relacionados. A Seção 3 descreve a metodologia adotada, desde a construção do *dataset* até o uso do *Chain of Thought* para classificação. A Seção 4 apresenta os resultados obtidos na avaliação dos modelos. A Seção 5 realiza as discussões, limitações do estudo e possibilidades para trabalhos futuros. Por fim, a Seção 6 apresenta as considerações finais e principais contribuições.

2. Fundamentação Teórica e Trabalhos Relacionados

Nesta seção, apresentam-se os principais conceitos e fundamentos utilizados neste estudo. Para tal, buscou-se referências em pesquisas da área de Inteligência Artificial aplicada à Educação e ao Processamento de Linguagem Natural. Abordam-se o uso de LLMs no contexto do ensino-aprendizagem e tecnologias emergentes voltadas à classificação automática de textos.

2.1. Tecnologias emergentes na Educação Básica

Nos últimos anos, o avanço das tecnologias de Inteligência Artificial (IA) trouxe novas perspectivas para o desenvolvimento de recursos educacionais. Estudos têm demonstrado que a IA pode apoiar tanto o processo de ensino-aprendizagem quanto a gestão acadêmica, atuando na predição da evasão escolar, na análise de desempenho, na personalização de materiais didáticos e no auxílio à elaboração de planos de aula.

No trabalho de [Maia and Sarkis 2025], o uso de LLMs, como ChatGPT[1] e DeepSeek[2], demonstrou um impacto positivo no ensino-aprendizagem de programação Python para iniciantes. No âmbito da tomada de decisões educacionais, a IA apresenta alta capacidade para a previsão e gestão do desempenho acadêmico e da evasão de estudantes, [Rodrigues et al. 2024] investigaram o uso de modelos *Transformer* para previsão de desempenho acadêmico no ensino fundamental e médio, permitindo a identificação precoce de alunos em risco de reprovação para intervenção e apoio adequados.

A pesquisa de [Laranjeira and Bezerra 2025] propõe um sistema web que utiliza LLMs para auxiliar professores da rede pública brasileira na criação de planos de aula sobre pensamento computacional e tecnologia, simplificando a engenharia de prompts e gerando recursos educacionais alinhados à BNCC, considerando o contexto cultural e socioeconômico dos alunos. Esse sistema visa reduzir a carga de trabalho dos professores e promover a integração interdisciplinar, mostrando-se uma ferramenta acessível e eficaz.

A introdução da arquitetura *Transformer* [Vaswani et al. 2017] abriu caminho para o desenvolvimento de modelos de linguagem de grande porte, que se tornaram referências no campo do Processamento de Linguagem Natural. Dentre eles, a família de modelos GPT (*Generative Pre-trained Transformer*), desenvolvida pela OpenAI[3]que consolidou LLMs autoregressivos para geração de texto, bem como o BERT (*Bidirectional Encoder Representations from Transformers*), proposto pelo Google [Devlin et al. 2019], que destacou-se em tarefas supervisionadas de classificação textual.

Nesta perspectiva, uma técnica que vem sendo amplamente utilizada em modelos *Transformers* é o *Chain-of-Thought Prompting* (CoT), ou Prompting de Cadeia de Pensa-

[1] https://chatgpt.com/
[2] https://www.deepseek.com/
[3] https://openai.com/

mento, que explora a geração de uma série de etapas de raciocínio intermediárias, melhorando a capacidade de LLMs em realizar raciocínios complexos. Essa técnica demonstra ganhos relevantes em tarefas que exigem lógica e decomposição de problemas, simplesmente por fornecer exemplos que incluem uma sequência de passos para a resolução da tarefa no prompt, sem a necessidade de *finetuning* [Wei et al. 2022].

O trabalho de [Zhang et al. 2023] explorou a aplicação da estratégia de prompting CoT no modelo GPT-3.5 para a tarefa de detecção de posicionamento (*stance detection*) em redes sociais, com o objetivo de identificar a atitude do usuário diante de um alvo específico (tópico ou entidade). Em outra linha de investigação, [Peres 2023] analisou o desempenho de LLMs com técnicas de *in-context learning*, como o CoT, na resolução de questões textuais complexas de vestibulares militares brasileiros (IME e ITA). O estudo avaliou modelos como text-davinci-003, GPT-3.5-turbo e GPT-4 em provas de Língua Portuguesa, Matemática, Química e Física, considerando apenas questões que não dependiam de imagens para sua resolução.

Os resultados dos estudos citados indicam que a técnica CoT atinge desempenho satisfatório em tarefas de classificação e resolução de problemas complexos [Zhang et al. 2023, Peres 2023]. Sob essa ótica, a utilização de LLMs com a técnica CoT apresenta-se como uma alternativa para a criação de aplicações com fins educacionais, contribuindo diretamente para a aprendizagem em Computação na Educação Básica.

3. Metodologia

A metodologia foi estruturada em três etapas: (i) mineração e extração de textos de questões da OBI; (ii) treinamento supervisionado com o modelo BERTimbau; e (iii) prompting com *Chain of Thought*. Essa abordagem permitiu comparar métodos tradicionais de aprendizado supervisionado com técnicas recentes de prompting, avaliando seu potencial para a classificação automática das questões da competição.

3.1. Mineração de Textos e Extração de Questões da OBI

O processo de desenvolvimento deste classificador teve início com a coleta e pré-processamento de dados textuais. Inicialmente, foi realizado o download dos cadernos de provas da OBI. Para automatizar esse processo, optou-se pela biblioteca *Selenium*[4] com Python, possibilitando a coleta de um conjunto de provas de 2003 a 2024 de diferentes fases e níveis, obtidas diretamente do site oficial [5] da competição.

Na sequência, os arquivos PDF foram convertidos para o formato *Markdown* (.MD) utilizando a biblioteca *LlamaParse*[6], que permitiu a extração de textos das provas e organização das questões, resultando em um conjunto de dados (*dataset*) no formato CSV/UTF-8 onde cada registro de questão foi estruturado com os seguintes atributos: ano, fase, nível, número_questão, título, enunciado, questão e alternativas. A padronização nesse formato permitiu a sistematização dos dados e sua integração em pipelines de análise e experimentação.

Após a extração dos textos, foram obtidas 2.568 questões. Em seguida, aplicaram-se filtros para selecionar apenas aquelas apresentadas integralmente em formato textual,

[4] https://selenium-python.readthedocs.io/
[5] https://olimpiada.ic.unicamp.br/
[6] https://pypi.org/project/llama-parse/

excluindo problemas que dependiam da análise de imagens ou elementos gráficos para sua resolução. Com isso, o conjunto foi reduzido para uma amostragem de 2.208 questões válidas.

Na etapa final do processo, utilizou-se uma planilha contendo questões rotuladas manualmente por um pesquisador doutor em Ciência da Computação, com atuação na área de Educação em Computação e experiência na análise de problemas da OBI, os textos extraídos do site oficial da competição foram organizados em um *dataset*[7] de referência composto por 412 questões.

A rotulagem foi conduzida com base em critérios definidos pelo pesquisador, considerando três categorias principais: (i) ordenação, envolvendo problemas relacionados à definição de ordem, posição ou arranjo de objetos; (ii) agrupamento, referente à atribuição de objetos a um ou mais grupos; e (iii) outros, contemplando questões que não se enquadram nas categorias anteriores, como cálculos, representações ou condições específicas [Martins 2011]. Esse processo de rotulagem e revisão resultou no *dataset* utilizado como base para os treinamentos e experimentos de classificação realizados neste trabalho.

3.2. Treinamento Supervisionado com BERTimbau

Na segunda etapa, foi conduzido o treinamento supervisionado de modelos de classificação de texto utilizando o *dataset* de referência. Observou-se que a distribuição das instâncias entre as classes era desbalanceada, com predominância de Ordenação (203 questões), seguida de Agrupamento (148 questões) e Outros (61 questões). Essa característica, está detalhada na Tabela 1.

Tabela 1. Questões Rotuladas OBI

Classe	Quantidade de Questões
Ordenação	203 - 49,3%
Agrupamento	148 - 35,9%
Outros	61 - 14,8%

Para a realização dos experimentos iniciais de classificação, optou-se pelo BERTimbau[8], devido ao modelo baseado no BERT, desenvolvido pela *NeuralMind*[9], ser pré-treinado em língua portuguesa, o que melhora seu desempenho em conjuntos de dados em português [Souza et al. 2020], como no caso das questões da OBI. Essa escolha se justifica pelo fato de sua arquitetura, fundamentada em transformers, possibilitar a captura de relações contextuais mais ricas nos textos [Devlin et al. 2019]. O objetivo consistiu em testar o desempenho do algoritmo na identificação automática de categoria das questões, com base nos rótulos definidos manualmente pelo especialista.

Devido à característica do *dataset* de referência com poucos dados rotulados e do desbalanceamento entre classes, para avaliação do classificador supervisionado, o modelo BERTimbau (neuralmind/bert-base-portuguese-cased) foi submetido a um processo de validação cruzada estratificada com K = 5 folds. Que consiste em particionar os dados de forma aleatória em cinco subconjuntos de tamanho igual, a cada iteração,

[7] https://tinyurl.com/ycxkknsz
[8] https://huggingface.co/neuralmind/bert-base-portuguese-cased
[9] https://neuralmind.ai/

uma delas é usada como conjunto de validação e as k - 1 demais são usadas para treino [Caseli and Nunes 2024]. Essa técnica permite estimar a capacidade de generalização do modelo, reduzindo o viés decorrente de uma única divisão treino/validação e fornecendo métricas mais confiáveis para conjuntos de dados de menor escala.

A configuração dos hiperparâmetros foi realizada de acordo com valores amplamente utilizados em experimentos já validados com o uso de Transformers, como sugerido por [Devlin et al. 2019]. Todo o processo foi conduzido no ambiente Google Colab[10], utilizando GPU NVIDIA Tesla T4 (16GB VRAM), para acelerar as operações. A Tabela 2, apresenta as configurações do modelo.

Tabela 2. Configuração dos parâmetros de treinamento

Parâmetro	**Valor**
Tokenizador	BertTokenizer
Tamanho Máximo da Sequência	512 tokens
Truncamento	True (longest_first)
Seed Global	42
Validação	5-fold estratificado
Batch Size	8
Épocas Máximas	2
Taxa de Aprendizado	2e-5
Otimizador	AdamW

Em complemento, foram consideradas métricas padrões, complementares à Acurácia e Precisão, como o Revocação e Medida-F Macro, por serem amplamente aceitas e recomendadas na literatura [Caseli and Nunes 2024]. Os resultados foram obtidos em relatórios de classificação do treinamento por meio da biblioteca Python de código aberto *Scikit-learn*[11].

3.3. Estratégia de Prompting com Chain of Thought

Considerando a importância de testar diferentes abordagens, além do treinamento supervisionado com BERTimbau, neste trabalho, realizou-se experimentos com um classificador automático baseado em CoT [Wei et al. 2022], com o objetivo de avaliar sua eficácia na categorização de questões da competição.

Para guiar o LLM no processo de classificação, foi construído um prompt do tipo *Few-Shot CoT*, que oferece alguns exemplos de entradas e saídas desejadas no próprio prompt, antes de solicitar a resposta para determinada tarefa [Nascimento 2024]. Dessa forma, o prompt[12] aplicado nesse trabalho foi estruturado a partir dos tipos de questões e das regras mais frequentes observadas nas provas da OBI: ordenação, agrupamento e outros.

Baseando-se nas definições de [Martins 2011], adotou-se a divisão do prompt em duas partes: a *System Message* (Figura 1), que estabelece o papel do modelo, instruindo-o a atuar como um especialista na classificação de questões da OBI e definindo as regras e

[10] https://colab.google/
[11] https://scikit-learn.org/stable/index.html
[12] https://tinyurl.com/2astuc45

limites da interação, e o *Prompt Template* (Figura 2), que apresenta exemplos e orienta o raciocínio passo a passo, finalizando com a solicitação de classificação da questão em uma das três categorias definidas.

System Message utilizado para classificação das questões:

Você é um especialista em classificar questões da OBI em três categorias específicas: ORDENAÇÃO, AGRUPAMENTO e OUTROS.
Siga este raciocínio passo a passo antes de classificar:
1. O objetivo principal da questão é definir ordem de objetos, posição ou arranjo ordenado? Isso inclui ordem explícita (1º, 2º, etc.) ou implícita (antes/depois, vizinhança, restrições de posição)?
Classifique como ORDENAÇÃO.
2. A questão foca na atribuição de objetos a um grupo? formação de subconjuntos ou seleção de elementos, sem considerar a ordem entre eles?
Classifique como AGRUPAMENTO.
3. A questão não se encaixa nas duas anteriores e envolve principalmente o cálculos de valores, análise de imagens, estruturas como grafos, tabelas ou algoritmos?
Classifique como OUTROS.

Finalize sempre com: Classificação Final: [ordenação—agrupamento—outros]

Figura 1. System message.

Prompt Template exemplos para classificação das questões:

EXEMPLOS DE CLASSIFICAÇÃO COM RACIOCÍNIO:

ORDENAÇÃO: Enunciado: Três funcionárias, Ana, Bia e Clara, trabalham em três andares diferentes de um prédio (1º, 2º e 3º). Questão: Se Ana não está no 3º andar e Clara não está no 2º, quais são as distribuições possíveis? Análise: A questão exige determinar a posição de cada pessoa. Isso é uma relação clara de ordem. Classificação Final: ordenação
AGRUPAMENTO: Enunciado: Para um combo de pizzas, o cliente escolhe 4 entre 7 sabores disponíveis. Questão: Qual das alternativas representa um grupo completo de sabores? Análise: O objetivo é formar um grupo. A ordem dos sabores não interfere na resposta. Classificação Final: agrupamento
OUTROS: Enunciado: Um torneio com 128 jogadores elimina um por rodada. Questão: Quantas rodadas são necessárias até restar um vencedor? Análise: Requer aplicação de fórmula matemática. Trata-se de cálculo. Classificação Final: outros

— Agora, classifique a seguinte questão:
Enunciado: enunciado Questão: questao
Siga o raciocínio passo a passo e analise o objetivo principal da questão.
Finalize com: Classificação Final: [ordenação—agrupamento—outros]

Figura 2. Prompt Template.

Esse formato de raciocínio explícito possibilitou a seguir uma sequência de passos lógicos antes de emitir a classificação final, favorecendo maior consistência no processo

decisório. O classificador CoT foi então aplicado ao *dataset* de 412 questões rotuladas manualmente, considerando apenas enunciados em formato textual. Os experimentos conduzidos com essa abordagem, utilizando as versões GPT-4.1-mini e GPT-5-mini da plataforma OpenAI, tiveram seu desempenho avaliado a partir de métricas de acurácia e de classificação, cujos resultados são apresentados na seção seguinte.

4. Resultados

Nesta seção, são apresentados os resultados e a avaliação de desempenho dos modelos propostos. Inicialmente, reportam-se as métricas do classificador supervisionado baseado no modelo BERTimbau. Em seguida, são detalhadas as métricas de classificação obtidas via *Few-Shot CoT*. Por fim, é realizada a comparação de seu desempenho com o modelo supervisionado.

4.1. Classificação Supervisionada com BERTimbau

Considerando a aplicação de validação cruzada estratificada com $K = 5$ folds, os resultados de desempenho do modelo BERTimbau (neuralmind/bert-base-portuguese-cased) na classificação supervisionada são detalhados na Tabela 3. O modelo demonstrou alto desempenho, alcançando uma acurácia média de 90,9% e uma Medida F Macro de 90,4%.

Tabela 3. Resumo do desempenho em validação cruzada com 5-folds

Modelo	Acurácia	Precisão	Revocação	Medida-F Macro
BERTimbau	0.9093	0.9457	0.8844	0.9048

Esses resultados indicam que o classificador conseguiu manter um equilíbrio adequado entre precisão e cobertura das classes, demonstrando boa capacidade de generalização mesmo em um conjunto de dados relativamente reduzido e desbalanceado. Embora o BERTimbau tenha alcançado bons resultados na tarefa de classificação supervisionada, essa abordagem depende fortemente de um *dataset* rotulado manualmente para garantir a qualidade da anotação [Britto et al. 2022].

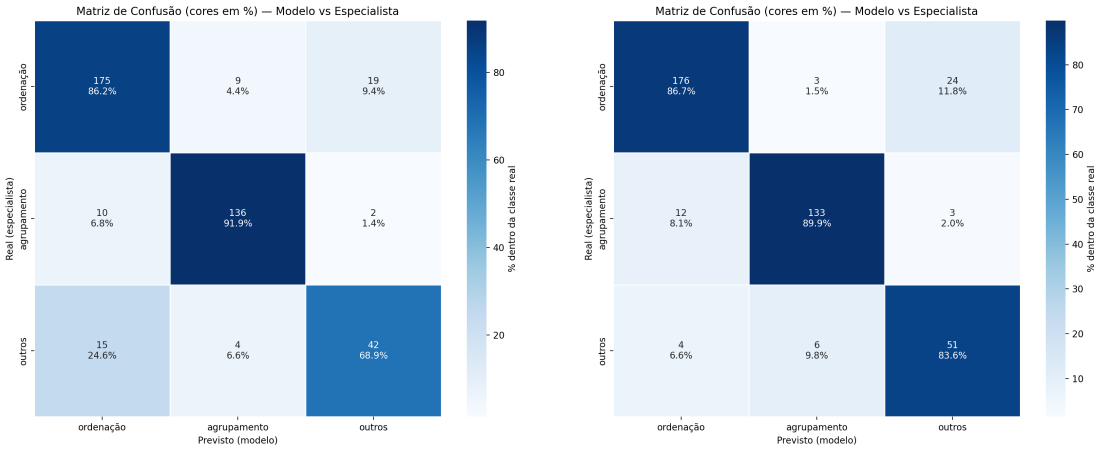
Para investigar alternativas mais leves em termos de anotação, a próxima subseção (4.2) apresenta os resultados obtidos com uma estratégia baseada em prompting com *Few-Shot CoT*, na qual o modelo de linguagem é guiado por exemplos de raciocínio passo a passo. Essa comparação busca avaliar em que medida a técnica pode se aproximar do desempenho supervisionado, ao mesmo tempo em que reduz a necessidade de rotulagem manual.

4.2. Classificação via Chain of Thought

Foram avaliados dois modelos disponibilizados pela plataforma OpenAI: o GPT-4.1-mini e o GPT-5-mini. A classificação via API da plataforma com *temperature=0* para a versão GPT-4.1-mini. Adotou-se a técnica *Few-Shot CoT* com o objetivo de reduzir a necessidade de bases rotuladas manualmente [Britto et al. 2022].

A Figura 3 apresenta as matrizes de confusão normalizadas dos dois modelos analisados. Em ambas, verifica-se predominância de acertos na diagonal correspondente às classificações corretas, evidenciando capacidade consistente de discriminação entre as classes.

No GPT-4.1-mini (Figura 3-a), a classe agrupamento apresentou maior taxa de acerto 91,9%, enquanto a classe outros demonstrou desempenho inferior 68,9%, com elevada confusão para ordenação 24,6%. Esse padrão sugere que problemas classificados como outros frequentemente apresentam elementos sequenciais ou condicionais que podem ser interpretados pelo modelo como indícios de ordenação.



(a) GPT-4.1-mini **(b) GPT-5-mini**

Figura 3. Matrizes de confusão normalizadas (%) dos modelos GPT-4.1-mini e GPT-5-mini na classificação das questões da OBI. As linhas representam as classes reais e as colunas as classes preditas.

No GPT-5-mini, observa-se na (Figura 3-b) redução da ambiguidade entre classes, com aumento da taxa de acerto na categoria outros para 83,6% e queda da classificação indevida como ordenação para 6,6%. Esses resultados indicam maior capacidade do modelo em distinguir entre problemas de natureza sequencial explícita e aqueles de caráter mais geral, reduzindo os erros na categorização.

A redução das ambiguidades pode estar relacionada às melhorias introduzidas na versão mais recente do modelo, no que se refere à capacidade de raciocínio estruturado e ao seguimento consistente de instruções, aspectos documentados nas atualizações da família GPT.

4.3. Comparação entre BERTimbau e Chain of Thought

A análise comparativa entre os modelos é apresentada na Tabela 4. Observam-se duas abordagens distintas para a classificação das questões. De um lado, o *BERTimbau*, que alcançou os melhores resultados em todas as métricas avaliadas, evidenciando a eficácia do aprendizado supervisionado quando há disponibilidade de dados rotulados. De outro, os modelos GPT-4.1-mini e GPT-5-mini, avaliados sob o mesmo conjunto de dados e as mesmas métricas, utilizando a estratégia de *Few-Shot CoT*.

A comparação entre os modelos baseados em LLM indica que o GPT-5-mini supera o GPT-4.1-mini em todas as métricas globais, com aumento na acurácia (87,4% vs. 85,6%), na precisão (83,5% vs. 81,8%), na revocação (86,7% vs. 82,3%) e na Medida-F Macro (84,7% vs. 82,0%). Esse avanço, ainda que incremental, indicam melhoria consistente no desempenho global e maior equilíbrio entre as três classes analisadas (ordenação, agrupamento e outros).

Ainda assim, o BERTimbau apresentou os melhores resultados gerais, alcançando acurácia de 90,9%, precisão de 94,6%, revocação de 88,4% e Medida-F Macro de 90,5%, superando os modelos baseados em LLM em todas as métricas consideradas.

Tabela 4. Comparação de desempenho dos modelos (GPT-4.1-mini, GPT-5-mini e BERTimbau)

Modelo	Acurácia	Precisão	Revocação	Medida-F Macro
GPT-4.1-mini	0.856	0.818	0.823	0.820
GPT-5-mini	0.874	0.835	0.867	0.847
BERTimbau	**0.909**	**0.946**	**0.884**	**0.905**

Diante das métricas analisadas, observa-se que a aplicação de *Chain of Thought* apresentou desempenho competitivo, com o GPT-5-mini se aproximando do BERTimbau nas métricas de acurácia e revocação, indicando maior estabilidade na identificação das classes.

5. Discussões

A literatura indica que diferentes estratégias podem ser empregadas na classificação textual. Conforme observado por [Rodrigues et al. 2024], algoritmos clássicos apresentam bom desempenho com menor custo computacional, mas dependem de etapas de pré-processamento. De modo semelhante, abordagens supervisionadas como o BERTimbau, utilizado neste estudo, alcançam alto desempenho quando há disponibilidade de dados rotulados. Em outra direção, [Peres 2023] evidenciou que modelos mais recentes, ao aplicarem a estratégia de CoT, obtêm melhor desempenho em tarefas complexas. No presente trabalho, essa tendência também foi observada, com desempenho superior do GPT-5-mini em relação à versão anterior.

As análises deste estudo indicam que a classificação automática de questões da OBI envolve desafios que vão além da categorização textual. A abordagem baseada em *Few-Shot CoT* apresentou resultados satisfatórios. Entretanto, o custo computacional e financeiro de modelos comerciais, como o GPT-4.1-mini e GPT-5-mini [Nascimento 2024], deve ser avaliado considerando as diretrizes e fomentos em educação digital previstos na PNED [Brasil 2023]. Esse cenário reforça a necessidade de equilibrar desempenho, viabilidade econômica e acessibilidade tecnológica.

Referente aos resultados de classificação obtidos neste estudo, segundo [Martins 2011], a relação entre a quantidade de objetos e a quantidade de posições ou elementos em grupos, definidos no enunciado da questão, pode influenciar no grau de dificuldade, questões com essas características são mais presentes a partir do nível 2 da competição. Ao obter o relatório de erros de classificação, constatou-se que a questão Reações Químicas (Figura 4), originalmente rotulada como Agrupamento, foi classificada como Ordenação pelos dois modelos GPT-4.1-mini e GPT-5-mini.

O autor define que ”regras do tipo condicional podem aparecer tanto em questões da categoria Ordenação quanto em questões da categoria Agrupamento”[Martins 2011, p. 27], pois devem ser selecionados um subconjunto dos objetos ou decidir que posições ou grupos que não serão preenchidos. Esse tipo de questão apresenta-se desafiadora até para os modelos de linguagens testados, visto que, a tarefa exige a atribuição de variáveis

Questão 19: Reações Químicas - OBI 2008 Fase 2 - Nível 2

Um químico está misturando substâncias para desenvolver dois novos produtos para a indústria. Para isso ele conta com nove tipos de substâncias: J, K, L, M, N, O, P, Q e R. Cada substância deve estar presente em apenas um dos dois produtos: o produto A ou o produto B, ou seja, se, por exemplo, J foi usado no produto A então J não pode ser usado no produto B. Exatamente cinco substâncias serão misturadas para obter um dos produtos e as outras quatro serão misturadas para obter o outro produto. As seguintes condições também se aplicam: - Se J está no produto a então tanto M quanto N devem estar no produto B. - no produto em que K estiver não pode estar nem M nem R. - a substância O deve ser a última substância a ser misturada no desenvolvimento do produto que contém exatamente quatro substâncias. - a substância P deve ser misturada ao produto que contém exatamente cinco substâncias. - a substância R deve ser a segunda a ser misturada no produto A ou deve ser a terceira a ser misturada no produto B. - a substância K deve ser a quarta a ser misturada independentemente do produto. - a substância L é a terceira a ser misturada no produto A.

19. Se Q é a terceira substância do produto B, qual das seguintes opções é uma lista correta e completa de substâncias que devem também ser misturadas no produto B?

Figura 4. Erro de Classificação - questão Reações Químicas

a dois grupos distintos, além de solicitar a ordem das variáveis nos respectivos grupos, caracterizando esse tipo de questão como Grupos Ordenados [Martins 2011]. Nesse caso, o prompt identificou ambiguidade entre o rótulo de referência e a classificação, demonstrando a complexidade da tarefa em cenários com múltiplas regras. Tais resultados indicam a necessidade de maior refinamento dos prompts e da definição de subclasses mais específicas, de modo a reduzir ambiguidades e melhorar a precisão da categorização.

Do ponto de vista pedagógico, ao classificar uma questão como agrupamento, ordenação ou outros, evidencia-se a mobilização de pilares fundamentais do Pensamento Computacional, conforme proposto por [Brackmann et al. 2019]. Problemas de ordenação envolvem raciocínio algorítmico e organização sequencial de variáveis, enquanto questões de agrupamento exigem abstração e reconhecimento de padrões, decompondo-os na formação de subconjuntos. Ambas as categorias analisam um cenário a partir da aplicação de regras condicionais. Assim, a identificação dessas categorias contribui para o desenvolvimento de habilidades cognitivas que facilitam a resolução de problemas de forma lógica e sistemática.

Como aplicação prática, a disponibilização de um *dataset* estruturado e de um classificador automático de questões, integrados a uma plataforma educacional, podem auxiliar estudantes e professores na organização de exercícios por categorias. Esse material contribui tanto para o treinamento da competição quanto para o planejamento de atividades voltadas ao desenvolvimento de competências previstas na BNCC Computação.

5.1. Limitações do estudo

O presente estudo apresenta algumas limitações, inicialmente, o tamanho reduzido do *dataset* com apenas 412 questões rotuladas manualmente, com desbalancemento entre classes, pode influenciar o desempenho dos classificadores. A decisão de testar e avaliar

um número limitado de modelos pré-treinados da OpenAI, outros LLMs recentes ou variantes especializadas da língua portuguesa podem apresentar desempenho superior, mas não foram explorados.

Não foram aplicadas abordagens de aprendizado semi-supervisionado ou de extração de padrões conceituais, o que abre espaço para futuras investigações sobre os pilares do PC e conceitos de Computação na Educação Básica presentes nas provas da competição.

5.2. Trabalhos futuros

Embora os resultados sejam promissores, este estudo abre novas perspectivas para pesquisas futuras, que podem expandir o escopo e a contribuição para a expansão da Computação na Educação Básica. Entre as principais direções, destacam-se:

- Expandir a categorização para outros tipos de questões da OBI, incluindo problemas de Cálculo, Grupos Ordenados e demais estruturas;
- Mapear os conceitos de Computação presentes nos enunciados, para identificar habilidades do PC e apoiar a curadoria de materiais educacionais;
- Integrar o classificador a plataformas de aprendizagem adaptativa, com filtragem de questões por categoria e geração de trilhas personalizadas;

6. Considerações finais

Este trabalho, em continuidade à linha de pesquisa apresentada em [de Almeida et al. 2024], resultou na construção de um classificador automático e de um *dataset*[13] inédito, estruturado com questões da Modalidade Iniciação da OBI. Como contribuição, o material organiza problemas de raciocínio lógico em categorias, ampliando as possibilidades de utilização pedagógica desse acervo e favorecendo o desenvolvimento de habilidades associadas ao Pensamento Computacional.

Os resultados obtidos indicam que o uso de LLMs com a técnica de *Chain of Thought* não se restringe à automação da classificação textual, podendo também apoiar a organização de materiais de treinamento voltados à preparação de estudantes para competições de lógica e informática. A categorização dessas questões possibilita a criação de trilhas de estudo, a seleção de problemas por tipo de raciocínio e a exploração dos desafios mais recorrentes na OBI.

Em contexto educacional mais amplo, o material também pode apoiar iniciativas de formação docente e pesquisas em Computação na Educação Básica. A organização das questões em categorias facilita sua utilização como recurso didático em sala de aula, permitindo que professores integrem desafios de lógica e resolução de problemas às atividades de ensino e promovam novas investigações na área.

7. Agradecimentos

O presente trabalho foi realizado com apoio da Fundação Universidade Federal de Mato Grosso do Sul (UFMS) e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

[13] https://tinyurl.com/yuv4kpb5

Uso de Inteligência Artificial

Durante a elaboração deste trabalho, foram utilizadas ferramentas de Inteligência Artificial Generativa como apoio em atividades específicas, incluindo a revisão e adequação de código da automação para mineração de textos (*Selenium e LlamaParse*), a realização de testes com o modelo BERTimbau e o suporte à revisão textual. Ressalta-se que, todo o processo foi revisado e validado pelos autores.

Referências

Brackmann, C. P., Caetano, S. V. N., and da Silva, A. R. (2019). Pensamento computacional desplugado: ensino e avaliação na educação primária brasileira. *Revista Novas Tecnologias na Educação*, 17(3):636–647.

Brasil (2022). CNE/CEB. Parecer Nº 2/2022 - Normas sobre Computação na Educação Básica – Complemento à BNCC. http://portal.mec.gov.br/index.php?option=com_docmanview=downloadalias=235511-pceb002-22category_slug=fevereiro-2022-pdfItemid=30192 - Acesso em: 26 ago. 2025.

Brasil (2023). Lei nº 14.533, de 11 de janeiro de 2023. institui a política nacional de educação digital. `https://www.planalto.gov.br/ccivil_03/_Ato2023-2026/2023/Lei/L14533.htm`. Acesso em: 26 ago. 2025.

Britto, L. F., Pessoa, L. A., and Agostinho, S. C. (2022). Cross-domain sentiment analysis in portuguese using bert. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 61–72. SBC.

Caseli, H. and Nunes, M. (2024). Processamento de linguagem natural: Conceitos, técnicas e aplicações em português. bpln, 2 edn.(2024).

Dagiene, V. and Stupuriene, G. (2016). Bebras–a sustainable community building model for the concept based learning of informatics and computational thinking. *Informatics in education*, 15(1):25–44.

de Almeida, T. G., Batista, E. J. S., de Lima, A. C., and Junior, A. A. C. (2024). Produçao e desenvolvimento de material de apoio ao treinamento para a modalidade iniciaçao da obi: Uma revisao sistemática da literatura. In *Workshop sobre Educação em Computação (WEI)*, pages 477–488. SBC.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Instituto de Computação - Unicamp (2025). Olimpíada brasileira de informática (obi). `https://olimpiada.ic.unicamp.br/`. Acesso em: 26 ago. de 2025.

Laranjeira, M. L. and Bezerra, P. T. (2025). Gerador automático de planos de aula sobre tecnologia baseado em llms. In *Workshop sobre Educação em Computação (WEI)*, pages 515–526. SBC.

Maia, S. M. and Sarkis, L. C. (2025). Utilização de llm como ferramenta de apoio no ensino-aprendizagem de programação python para iniciantes: Um relato de experiência. In *Workshop sobre Educação em Computação (WEI)*, pages 385–396. SBC.

Martins, W. S. (2011). *Jogos de Lógica: divirta-se e prepare-se para a Olimpíada Brasileira de Informática*. Vieira.

Nascimento, D. B. d. S. d. (2024). *Classificação automática de avaliações de acessibilidade em lojas de aplicativos: um estudo sobre técnicas de prompt*. PhD thesis, Universidade de São Paulo.

Peres, R. S. (2023). Grandes modelos de linguagem na resolução de questões de vestibular: o caso dos institutos militares brasileiros. Master's thesis.

Rodrigues, L. S., Santos, M., Gomes, C. F. S., Choren, R., Goldschmidt, R., and Barbará, S. (2024). Transformers para previsão de desempenho acadêmico no ensino fundamental e médio. *Revista Brasileira de Informática na Educação*, 32:213–241.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wing, J. (2016). Pensamento computacional–um conjunto de atitudes e habilidades que todos, não só cientistas da computação, ficaram ansiosos para aprender e usar. *Revista Brasileira de Ensino de Ciência e Tecnologia*, 9(2).

Zhang, B., Fu, X., Ding, D., Huang, H., Dai, G., Yin, N., Li, Y., and Jing, L. (2023). Investigating chain-of-thought with chatgpt for stance detection on social media. *arXiv preprint arXiv:2304.03087*.