Utilização de *Chain-of-Thought Prompting* para classificação de questões da OBI

Davi Queiroz Rodrigues[1], **Rodrigo Seiti Koga Kikuta**[1], **Amaury Antonio de Castro Júnior**[1]

[1]Faculdade de Computação – Universidade Federal do Mato Grosso do Sul (UFMS)
79070-900 – Campo Grande – MS – Brasil

davi_queiroz@ufms.br, rodrigo_seiti@ufms.br, amaury.junior@ufms.br

Abstract. *The Brazilian Informatics Olympiad (OBI), in its initiation category, evaluates the logical problem-solving abilities of basic education students; however, the study material recommended by the organizers is based on older editions of the competition, which may compromise preparation for more recent exams. This work investigates whether the classes and types defined in the recommended material remain adequate for current editions of the OBI, and analyzes how modern prompting techniques, especially Chain-of-Thought (CoT), can support the automatic classification of these questions. The methodology involved an evaluation of the exams and their questions, followed by the incremental development of different prompts applied to the GPT-4o model, starting from zero-shot and one-shot approaches and progressing to versions with explicit reasoning. The results indicate that the categories proposed in the study material remain suitable for recent exams and highlight the potential of CoT prompting as a support tool for educational analyses and the development of automated question classification systems. The main contribution of this work is the documentation of a prompt-writing methodology that can serve as a reference for future classification research involving Large Language Models (LLM).*

Resumo. *A Olimpíada Brasileira de Informática (OBI), em sua modalidade iniciação, avalia a capacidade de resolução de problemas de lógica de estudantes do ensino básico; contudo, o material recomendado pelos organizadores contemplam edições antigas da competição, o que pode comprometer a preparação para as provas mais recentes. Este trabalho investiga se as classes e tipos definidas no material recomendado permanecem adequadas às edições atuais da OBI, além de analisar com técnicas modernas de prompting, em especial o Chain-of-Thought(CoT), podem auxiliar na classificação automática dessas questões. A metodologia envolveu uma avaliação das provas e questões, seguida da escrita de diferentes prompts de forma incremental aplicados ao modelo GPT-4o, partindo de abordagens zero-shot e one-shot até versões com raciocínio explícito. Os resultados indicam que as categorias propostas no material de estudo continuam adequadas às provas mais recentes e evidenciam o potencial do CoT prompting como ferramenta de apoio para análises educacionais e desenvolvimento de sistemas de classificação automática de questões. A principal contribuição deste trabalho é a documentação de uma metodologia de escrita de prompt que pode ser utilizada como referência para futuros trabalhos de classificação com modelos de linguagem de larga escala (LLM).*

1. Introdução

A OBI [IC-UNICAMP 2000] é uma iniciativa de grande importância educacional, pois estimula o desenvolvimento do raciocínio lógico e a habilidade de resolver problemas entre os alunos do ensino básico e médio, além de incentivar a compreensão e a aplicação de conceitos essenciais da ciência da computação. A competição incentiva o pensamento organizado, a análise de padrões e a criação de soluções inovadoras por meio de desafios que se tornam progressivamente mais complexos, ajudando a desenvolver habilidades cognitivas fundamentais em um ambiente tecnológico que exige cada vez mais. Além disso, a OBI desempenha o papel de catalisadora do interesse em carreiras no campo da tecnologia da informação, servindo como ponto de partida para futuros profissionais e pesquisadores, ao passo que proporciona acesso a oportunidades acadêmicas, bolsas de estudo e envolvimento em competições internacionais, como a *International Olympiad in Informatics* (*IOI*).

A competição é organizada nos moldes das outras olimpíadas científicas brasileiras, é divida em duas modalidades: Iniciação e Programação. Na modalidade iniciação, alunos que ainda não sabem programar competem resolvendo problemas de lógica, conceitos de computação e matemática. A prova é dividida em níveis (júnior, 1 e 2) e fases (2 definidas pelo calendário da OBI, mas podem possuir mais), o nível júnior é voltado para alunos do 4º e 5º anos do ensino fundamental, o nível 1, para alunos do 6º e 7º anos do ensino fundamental e o nível 2, para alunos do 8º e 9º anos do fundamental. Como o foco da modalidade iniciação são os alunos do ensino básico o site da olimpíada oferece a ementa e uma seção de preparação, para alunos e professores que desejam encontrar materiais e guias de preparação para prova; porém o único material indicado nas páginas para a preparação da prova é o livro **Jogos de Lógica: divirta-se e prepare-se para a Olimpíada Brasileira de Informática** [Martins 2011].

Em seu livro Martins (2011, p. 13), apresenta uma classificação das questões da OBI, divididas em classes e tipos, e a partir da classificação propõe métodos de resolução gerais com a utilização de diagramação dos problemas. O livro utiliza como *dataset* para a criação de classes e tipos as provas aplicadas nas edições da OBI de 2003 a 2009.

No contexto educacional, o emprego de *LLMs* tem sido cada vez mais discutido para as mais diversas atividades, partindo de atividades relacionadas a geração de conteúdo [Al Faraby et al. 2024] a personalização do processo de aprendizagem em diferentes níveis de educação [Kasneci et al. 2023].

Diante desse contexto, esse artigo aborda uma verificação da adequação das classes e tipos apresentadas no livro Jogos de Lógica [Martins 2011], principalmente em relação as edições mais recentes, a escrita de um *prompt* que utiliza a técnica *CoT* para classificação de questões da OBI, para que sirva de complemento na preparação e treinamento para a realização da prova, e a apresentação dos resultados encontrados ao compararmos o uso de outras técnicas de *prompting*, em relação ao *CoT*. [Phoenix and Taylor 2024]

Os resultados obtidos demonstram que o material disponível mantém-se preciso quanto a classificação das questões da prova e indicam que a utilização de *LLMs* é viável para o processo de classificação automatizada dessas questões, em especial quando combinada com o *CoT*. Este artigo detalha todas as etapas de desenvolvimento do trabalho,

os desafios enfrentados, as soluções aplicadas e atividades futuras, como a utilização de *propmpts* para classificação de outras olimpíadas do conhecimento. Assim, busca-se contribuir para o avanço da utilização de *LLMs* no contexto educacional.

2. Trabalhos Relacionados

Na literatura atual existem diversas linhas de estudo do emprego de tecnologia da informação nos contextos da OBI e da utilização de *LLMs* na resolução de provas, mas ainda existem lacunas se o objetivo for classificar sistematicamente as questões segundo uma taxonomia. A seguir, dois trabalhos que dialogam diretamente com as linhas de pesquisa deste estudo, um para preparo da OBI e outro para o uso de *LLMs* na solução de provas.

O Artigo ***Logic in a Logic Way*: um Aplicativo para Exercitar a Resolução de Problemas de Lógica da Olimpíada Brasileira de Informática** [Trindade et al. 2017] mostra o desenvolvimento de um aplicativo de *smartphone* voltado para a modalidade Iniciação da OBI, que visa ajudar os estudantes no trabalho de resolução de problemas de lógica. Os autores refletem que a preparação para a OBI, em especial a etapa de raciocínio lógico textual, não possui recursos acessíveis, além do livro de referência e das provas anteriores. O aplicativo fornece uma metodologia sistemática de treino e ilustra que materiais de apoio no tratamento do raciocínio lógico podem promover a difusão e o treinamento dos competidores.

Já a dissertação de mestrado **Um estudo da aplicação de Grandes Modelos de Linguagem na resolução de questões de vestibular: o caso dos Institutos Militares brasileiros**[Peres 2023] investiga a capacidade de *LLMs* ao resolver questões complexas de vestibulares das instituições militares brasileiras (IME e ITA). Os modelos *text-davinci-003*, *GPT-3.5-turbo* e o *GPT-4* foram testados por meio de técnicas de *in-context learning*, e com isso, observou-se que os modelos mais avançados, em especial, aqueles que utilizam técnicas de *CoT*, tiveram um ganho de desempenho. Este estudo revela o potencial de *LLMs* em resolver questões que envolvem altos níveis de complexidade, tanto textual quanto lógico, mostrando algumas aplicações no ambiente educacional e de avaliação. No entanto, seu foco principal é com a resolução dessas questões.

3. Fundamento Teórico

Modelos de linguagem de grande escala, como o *GPT-4o*, são modelos que usam aprendizado de máquina para processar e analisar linguagem humana [OpenAI 2022]. Com a constante melhoria do seu algoritmo e poder de processamento, *LLM* têm a capacidade de processar e analisar dados contextuais, permitindo assim aplicações específicas. A precisão das respostas de uma *LLM* pode ser influenciada pelo uso de técnicas de *prompting* especializadas como, por exemplo, *zero-shot prompting*, *one-shot prompting* ou *CoT prompting* [Wolff 2023].

Segundo Phoenix e Taylor (2024, p.36), a construção eficaz de *prompts* segue cinco princípios fundamentais: dar direção, especificar formato, fornecer exemplos, avaliar qualidade e dividir o trabalho. Esses princípios atuam como um guia prático para converter uma atividade em instruções que o modelo possa entender:

- **Dar direção:** Assegura que o modelo compreenda o objetivo da tarefa e evite respostas genéricas;

- **Especificar formato:** Estabelece de forma clara como a saída deve ser organizada, diminuindo ambiguidades;
- **Providenciar exemplo:** Funciona como um método de aprendizado por demonstração, simplificando a reprodução de padrões desejados;
- **Avaliar qualidade:** Fortalece a natureza iterativa do processo, possibilitando a adaptação do *prompt* com base nos resultados alcançados;
- **Dividir o trabalho:** Auxilia no gerenciamento de tarefas complexas por meio de etapas menores que favorecem o raciocínio lógico.

Esses princípios oferecem um alicerce teórico robusto para a aplicação de técnicas como *zero-shot, one-shot* e *CoT prompting*, abordadas neste estudo, além de justificar a progressão metodológica adotada.

3.1. *Zero-Shot Prompting*

Zero-shot prompting é a técnica na qual é requisitado ao modelo que faça uma tarefa sem treinamento prévio ou exemplos daquela tarefa em específico [Ramlochan 2023], utilizando apenas o seu conhecimento pré-existente para inferir como lidar com essa nova tarefa. A técnica é boa para tarefas genéricas mas é imprecisa quando lida com atividades com muita complexidade.

3.2. *One-Shot Prompting*

Essa técnica disponibiliza ao modelo um exemplo para guiar a resolução da tarefa antes de apresentar a tarefa em si [Wolff 2023]. Mostrando um problema relacionado e a sua solução, o modelo agora tem uma base para resolução da tarefa. Apresenta resultados mais precisos que o do *zero-shot*, mas ainda assim pode apresentar erros pela falta de entendimento contextual.

3.3. *Chain-of-Thought Prompting*

CoT prompting é a técnica que força uma *LLM* executar a tarefa com um foco maior no contexto. As *LLMs* funcionam prevendo a próxima palavra em uma sequência baseado no contexto das palavras anteriores, isso acaba por vezes causando a perda da semântica da análise ou conversa. Ao utilizarmos o *CoT*, a tarefa é dividida em passos lógicos, de forma que possamos guiar a *LLM* a uma linha de pensamento que se assemelha a como um humano resolveria determinada tarefa. [Wei et al. 2022]

A implementação da técnica geralmente inclui um exemplo da tarefa com um conjunto de decisões tomadas para determinada conclusão e um *prompt* que deixa explícito a necessidade de demonstrar os passos tomados para a decisão. Por exemplo, em um problema matemático, pedir para que o modelo demonstre cada passo efetuado do cálculo. Isso força o modelo a atacar o problema de forma sequencial, bem como um humano faria para um problema extenso ou complexo.

4. Classificação das Questões

Como abordado no livro Jogos de Lógicas [Martins 2011], as questões podem ser classificadas em ”Classes”, que determinam a utilização das variáveis apresentadas na questão, e cada classe possui um conjunto de ”Tipos”que determinam o tipo de restrição que será aplicada as variáveis da questão.

4.1. Ordenação

A classe ordenação envolve posicionar elementos em uma sequência específica relativa a algum sistema. Os tipos dessa classe são:

- **Linear:** Lidam com uma única estrutura de diagrama linear;
- **Quadrática:** Lidam com uma ou mais estruturas de diagrama lineares sobrepostas;
- **Circular:** Lidam com uma estrutura de diagrama em que as extremidades se encontram;
- **Livre:** Lidam com uma estrutura de diagrama não definida ou que se assemelha a um grafo.

4.2. Agrupamento

A classe agrupamento divide elementos em grupos com base em características ou condições específicas. Os tipos dessa classe são:

- **1 Grupo:** Elementos pertencem a um único grupo com regras condicionais;
- **N-Grupos:** Elementos são divididos em dois ou mais grupos, frequentemente envolvendo regras de combinação e posição.

4.3. Outros

A classe outros inclui questões não contempladas nas classes anteriores, que podem combinar elementos de ordenação e agrupamento, ou que requerem cálculos matemáticos ou leitura e resolução de diagramas. Os tipos dessa classe são:

- **Grupos Ordenados:** Combinam as propriedades de Ordenação com as de Agrupamento;
- **Cálculo:** Dependem exclusivamente de um cálculo envolvendo as variáveis do "Cenário";
- **Definição (Dedução):** Apresentam algum conjunto de regras lógicas que devem ser seguidas ou algum conceito para encontrar a solução.

5. Metodologia

Para assegurar que os *prompts* escritos fossem eficazes no auxílio da classificação das questões, foram seguidos os passos conforme a figura 1 abaixo:

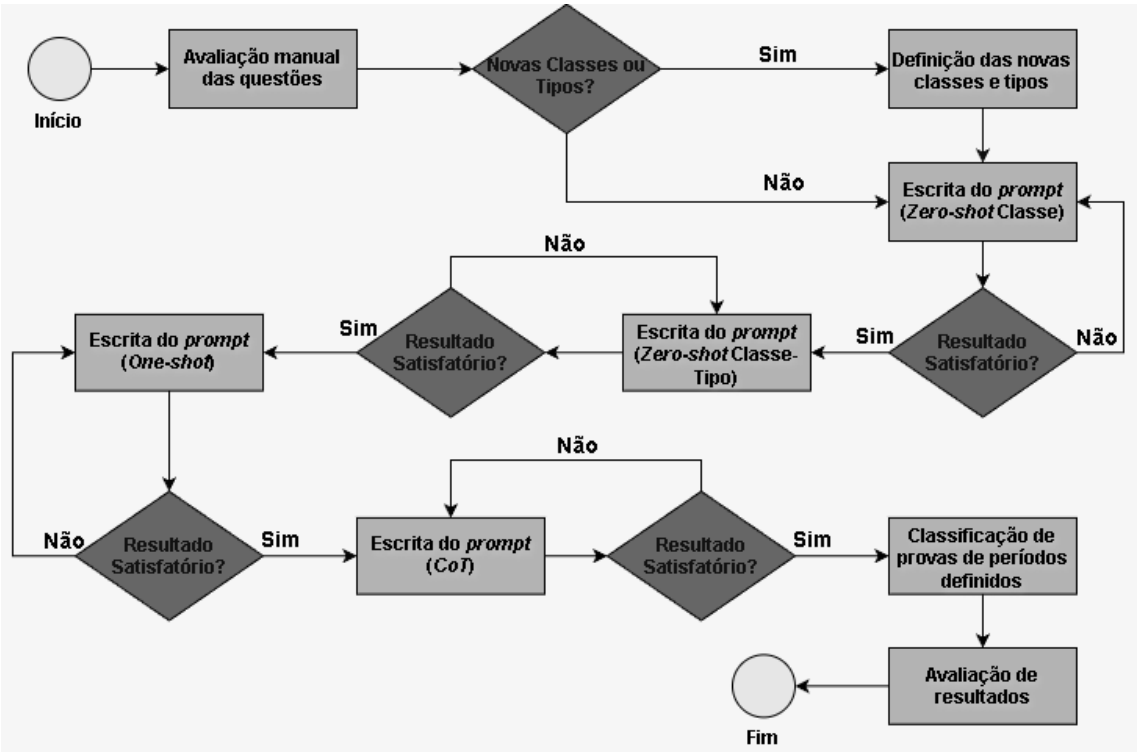

Figura 1. Diagrama de fluxo da metodologia

5.1. Avaliação manual do *dataset*

Inicialmente foi realizada uma revisão e classificação manual de todas as questões disponíveis até 2024, totalizando em 2789 questões classificadas manualmente, com o intuito de entender os critérios de classificação, verificar alterações na estrutura das questões ao longo das edições e avaliar a necessidade de alguma classe ou tipo novo para a classificação.

5.1.1. Estrutura das questões

Segundo Martins (2011, p. 23), uma questão típica é composta por três partes: Cenários, Regras e Perguntas; durante a revisão inicial das questões foi observado que ao longo das aplicações da OBI pouco foi alterado em relação a composição das questões, então aprofundando a definição inicial de Martins (2011, p. 28) temos:

- **Cenário:** Apresenta uma história que serve de contexto para pergunta, contém as variáveis que serão utilizadas na pergunta (locais, eventos, objetos, etc). Existem dois tipos que devem ser levados em consideração ao escrever um *prompt* para classificar uma questão: Cenários Compartilhados e Cenários Únicos:
  - Cenários Compartilhados: Após a descrição das regras esse cenário será utilizado por mais de uma pergunta. Em cenários compartilhados a questão sempre possuirá um título do cenário.
  - Cenários Únicos: A regra está, completamente ou parcialmente, contida no cenário e apenas uma única pergunta utilizará esse cenário. Em cenários únicos a questão pode ou não conter um título.

- **Regras:** Um conjunto de proposições que definem relações entre as variáveis (existência, posição, valor verdade).
  - Nas provas mais recentes aumentou-se o uso de recursos visuais para representação de regras. Mas a utilização não está limitada as regras, podendo aparecer no cenário ou até mesmo como alternativa de uma determinada pergunta.
- **Perguntas:** Conjunto de perguntas relacionadas ao cenário e às regras:
  - Em provas anteriores ao ano de 2012, as perguntas são demarcadas apenas por um número (1., 2., 3., . . .), já as demais são demarcadas por "Questão" seguido pelo número da pergunta, tudo em negrito.
  - Em perguntas de cenário compartilhado, é comum perguntas que alteram regras para o escopo da pergunta.

5.2. Utilização das classes e tipos

Durante a classificação manual alguns candidatos novos de tipos foram contemplados, porém como a quantidade de questões que se enquadravam nesses novos tipos não eram expressivas, e essas questões poderiam ser mantidas em classes e tipos previamente estabelecidas no livro [Martins 2011], decidimos seguir para a escrita dos *prompts* utilizando as classes e tipos apresentadas na Seção 4.

5.3. Escrita do *prompt zero-shot* Classes

Uma vez definido a estrutura base de uma questão da OBI e o escopo da análise, que compreendia todas as provas nos período de 2003 a 2009 e 2023 a 2024, escrevemos um *prompt zero-shot* capaz de reconhecer as diferentes partes de uma questão e avaliar apenas a classe da questão com base nas partes reconhecidas.

Pensando nos princípios apresentado por Phoenix e Taylor (2024, p.36), o *prompt* foi dividído em três módulos, o primeiro lidava com com a contextualização da OBI e realizava a atividade de identificação dos cenários, regras e perguntas dentro do arquivo *PDF*; o segundo lidava a classificação, nele é inserido a estratégia que o *prompt* utilizará, bem como as classes e tipos; por fim, o terceiro módulo especifica como o modelo deve apresentar o resultado, definimos duas formas, a primeira ao classificar a prova por inteira, onde o modelo devolverá um arquivo *CSV*, contendo o identificador de cada questão e a classificação atribuída, e a segunda para classificar uma única questão em específica, onde o modelo devolverá texto simples, contendo o identificador da questão, resumo do contexto dela, classificação atribuída e motivação da classificação aplicada. A classificação foi utilizada como um apoio para entendermos a classificação errônea do modelo e identificarmos pontos de melhoria para os próximos ciclos da iteração.

O *prompt* foi executado em um ambiente de memória temporária utilizando o modelo *GPT-4o*, a fim de não interferir com resultados posteriores. O resultado foi registrado e comparado com a análise manual.

A figura 2 apresenta o segundo módulo do *prompt zero-shot* com apenas classes, o *prompt* continha uma explicação dos itens que consistiam uma questão e uma breve descrição da definição das classes.

As ”Questões”da prova da OBI podem ser classificadas da seguinte maneira:
Ordenação: Envolve posicionar elementos em uma sequência específica.
Agrupamento: Divide elementos em grupos com base em características
ou condições específicas.
Outros: Inclui questões que combinam elementos de ordenação e agrupamento
ou que requerem cálculos matemáticos.

Figura 2. Excerto do *prompt Zero-Shot* (Classe)

5.4. Aplicação dos tipos no *prompt* inicial

Após a comparação com a análise manual, o *prompt zero-shot* foi modificado, como demonstrado na figura 3, recebendo um exemplo de questão para identificação dos componentes envolvidos em uma questão e uma descrição dos tipos associadas em suas classes. Um novo ambiente de memória temporária foi utilizado e o resultado foi registrado e comparado com a análise manual e a última versão testada.

Nos resultados desse *prompt*, em relação as provas mais recentes, foi notado que a presença de muitos recursos visuais atrapalhava na classificação de questões, então passamos a buscar alternativas para implementar nos próximos ciclos, a fim de evitar esse problema.

As questões podem ser classificadas em “Classes”, que determinam a utilização das variáveis apresentadas nos “Cenários”, cada classe possui um conjunto de “Tipos” que especificam o tratamento das “Regras” e variáveis do “Cenário”.

Ordenação: Envolve posicionar elementos em uma sequência específica relativa a algum sistema. Os tipos dessa classe são:
- Linear: Lidam com uma única estrutura linear;
- Quadrática: Lidam com uma ou mais estruturas lineares sobrepostas;
- Circular: Lidam com uma estrutura em que as extremidades se encontram;
- Livre: Lidam com uma estrutura não definida ou que se assemelha a um grafo.

Figura 3. Excerto do *prompt Zero-Shot* (Classe-Tipo)

5.5. Escrita do *prompt one-shot*

Após a comparação dos resultados obtidos até então a próxima etapa foi a inserção de exemplos de questões classificadas em cada tipo no *prompt*, apresentado na figura 4. Foi realizada uma escolha cuidadosa no exemplo escolhido em cada tipo para que não acontecesse ambiguidade no entendimento das classificações. Os dados foram coletados e o resultado foi comparado com os resultados dos passos anteriores.

Para o auxílio de questões onde as regras eram apenas recursos visuais introduzimos a utilização de palavras-chaves, de forma que a classificação pudesse ocorrer mesmo apenas com o texto do cenário.

Ordenação: Envolve posicionar elementos em uma sequência específica relativa a algum sistema.
É possível encontrar no cenário de questões dessa classe as palavras-chaves: *listar*, *classificar*, *sequência*.
Os tipos dessa classe são:
- Linear: Lidam com uma única estrutura linear; Não possui palavras-chaves, mas apresentam até dois conjuntos de variáveis que devem ser considerados para a ordenação.

Exemplo de Questão: "Empresas de busca na internet, como Bing e Google, classificam as páginas da Internet de acordo com a sua 'popularidade'. A popularidade de uma página X pode ser medida por exemplo pelo número de referências (links) de todas as outras páginas para X. Estamos interessados em seis páginas – P, Q, R, S, T e U,
que têm popularidades diferentes entre si. As seguintes relações são conhecidas:
P é mais popular do que Q ou R, mas não mais popular do que ambas.
U é menos popular do que R.
Se Q é menos popular do que R, então nem S nem U são mais populares do que T.
Se Q é mais popular do que R, então S é mais popular do que ambas T e U."

Figura 4. Excerto do *prompt One-Shot*

5.6. Utilização do *chain-of-thought prompting*

Por fim, foi escrita a versão do *prompt* contendo *CoT*, a figura 5 demonstra a adição da linha de pensamento de classificação dos exemplo de ordenação linear. Os dados foram comparados com os outros até então obtidos.

O raciocínio apresentado no *CoT*, envolvia a utilização das variáveis contidas no cenário, o contexto de como os elementos deviam ser tratados e as relações entre elas nas regras, e em como a partir dessas informações poderíamos inferir na classificação. A única exceção foi o tipo definição, onde o raciocínio de classificação é baseado na ausência de tratamento de elementos e regras que apresentavam especificações de contexto do cenário.

Ordenação: Envolve posicionar elementos em uma sequência específica relativa a algum sistema.
É possível encontrar no cenário de questões dessa classe as palavras-chaves: *listar*, *classificar*, *sequência*.
Os tipos dessa classe são:
- Linear: Lidam com uma única estrutura linear; Não possui palavras-chaves, mas apresentam até dois conjuntos de variáveis que devem ser considerados para a ordenação.

Exemplo de Questão: "Empresas de busca na internet, como Bing e Google, classificam as páginas da Internet de acordo com a sua 'popularidade'. A popularidade de uma página X pode ser medida por exemplo pelo número de referências (links) de todas as outras páginas para X. Estamos interessados em seis páginas – P, Q, R, S, T e U –, que têm popularidades diferentes entre si. As seguintes relações são conhecidas:
P é mais popular do que Q ou R, mas não mais popular do que ambas.
U é menos popular do que R.
Se Q é menos popular do que R, então nem S nem U são mais populares do que T.
Se Q é mais popular do que R, então S é mais popular do que ambas T e U."

Entrada: Analise "Cenário"e "Regras"acima, e classifique a questão.
Resposta: No caso do Exemplo dado é necessário fazer uma relação de popularidade de seis páginas(P, Q, R, S, T e U), a popularidade é medida pela incidência de cada página nas demais, todas com quantidades diferentes, assim estamos lidando com uma única estrutura linear, portanto a classificação é "Ordenação - Linear".

Figura 5. Excerto do *prompt CoT*

Uma das preocupações recorrentes durante o desenvolvimento desse *prompt* foi o tamanho final do comando enviado ao modelo, já que o uso do *CoT* tende a aumentar significativamente o número de *tokens* processados. Quanto maior o *prompt*, maior o custo computacional e maior o risco de perda de precisão, seja ou por extrapolar limites de contexto, ou por introduzir ruído desnecessário. Dessa forma, buscou-se um equilíbrio entre fornecer um bom raciocínio para orientar o modelo e evitar a criação de um *prompt* extenso, garantindo um processo de raciocínio detalhado sem comprometer a eficiência ou a interpretação correta das etapas da classificação.

6. Resultados

6.1. Desempenho dos *prompts*

No total 899 questões foram classificadas pelos *prompts*, sendo 449 do período de 2003 à 2009 e 450 do período de 2023 à 2024, demonstrando como a OBI mudou ao longo dos anos no formato e quantidade de provas por edição.

A Tabela 1, contém a quantidade de questões por tipo nos dois períodos classificados manualmente, podemos ver que no primeiro período a distribuição dos tipos de questão era mais uniforme, enquanto no segundo questões do tipo ordenação linear, cálculo e definição eram mais frequentes.

Tabela 1. Distribuição dos tipos de questões por período.

Tipo	2003–2009	%	2023–2024	%
Ord. – Linear	88	19,6	155	34,4
Ord. – Quadrática	64	14,3	19	4,2
Ord. – Circular	13	2,9	12	4,8
Ord. - Livre	38	8,5	3	0,7
Agrp. – 1 Grupo	55	12,2	8	1,8
Agrp. – N-Grupos	98	21,8	22	4,9
Cálculo	33	7,3	111	24,7
Grupos Ordenados	15	3,3	43	9,6
Definição	45	10,0	75	16,7
Total	449	100	450	100

As figuras 6 e 7, referentes aos períodos de 2003 a 2009 e 2023 a 2024, respectivamente, apresentam a taxa de acertos, erros e acertos parciais, por tipo de estratégia de *prompt* utilizada. No *prompt zero-shot* inicial, como o intuito era a identificação da estrutura das questões não utilizamos a classificação de tipos, portanto apenas após a segunda iteração do *prompt* consideramos os acertos parciais, onde o modelo classificava corretamente a classe porém errava o tipo. A ocorrência desse acerto parcial, na maioria das vezes, ocorria em questões na qual a interpretação das variáveis envolvidas podia ser aplicada em dois tipos distintos. Por exemplo, uma questão que envolvesse um grafo contendo um ciclo com regras de ordenação, poderia levar o modelo a classificar tanto como ordenação livre como ordenação circular.

Um ponto importante a considerar é a presença de imagens na questão, questões que são parcialmente ou completamente dependentes de imagens são na maioria das vezes ou classificadas como ”outros - definição”ou classificadas de forma errada, já que o modelo não lida tão bem com o contexto apresentado em formato de imagem, mesmo se tratando de um modelo multimodal; ainda assim na figura 7, o *prompt zero-shot* classe-tipo, mesmo tendo problemas com as imagens, teve uma redução na taxa de erros em relação ao *prompt zero-shot* apenas classe.

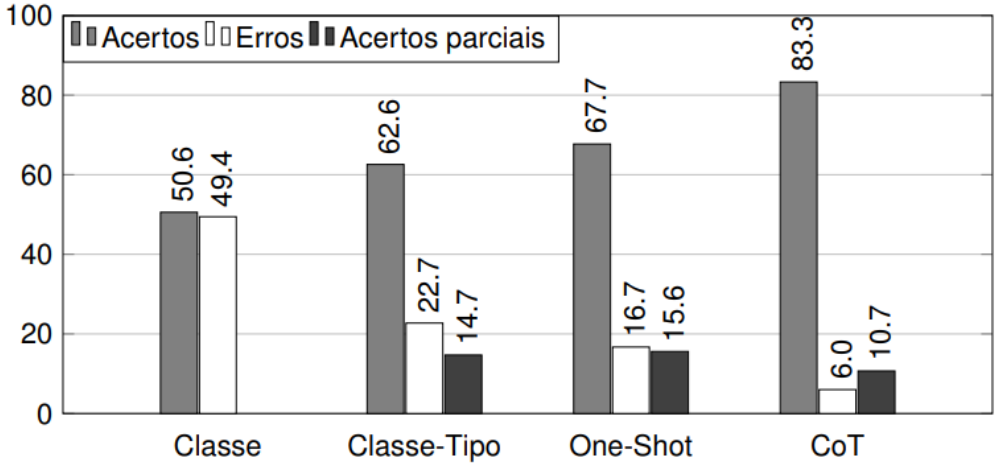

Figura 6. Gráfico de resultado da classificação das provas de 2003 a 2009

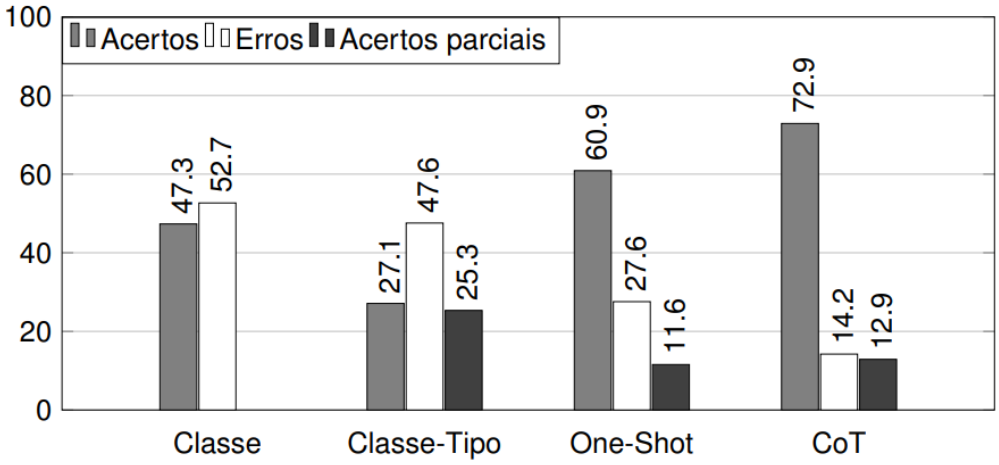

Figura 7. Gráfico de resultado da classificação das provas de 2023 a 2024

6.2. Avaliação das Classes e Tipos

Apesar das mudanças ao longo das edições da OBI, tanto no número de questões, fases de prova e níveis de escolaridade avaliados as classes e tipos apresentados no livro [Martins 2011], continuam sendo adequadas para classificação precisa das questões, sem que haja necessidade de modificarmos ou expandirmos as definições apresentadas por Martins (2011, p.28).

7. Ameaças na validade do estudo

Apesar de os resultados sugerirem que o uso de *CoT prompting* para classificar automaticamente as questões da OBI é viável, algumas ameaças à validade devem ser levadas em conta ao interpretar os percentuais apresentados. A primeira refere-se à inclusão de imagens nos exames mais recentes, que incorporam elementos visuais que o modelo não consegue captar completamente. Como o processo de classificação foi realizado principalmente com base em texto, perguntas que exigem diagramas, tabelas ou representações gráficas para serem compreendidas podem ter sido classificadas de maneira inadequada, o que diminui a precisão em relação à classificação manual.

Ademais, a presença de questões semanticamente similares, particularmente aquelas que compartilham o mesmo contexto com pequenas alterações nas regras ou no enunciado, constitui outra ameaça significativa. Em certos casos, o modelo pode interpretar essas questões como distintas, resultando em uma classificação inconsistente.

Somam-se a esses fatores potenciais limitações relacionadas ao tamanho do *prompt*, pois a inclusão de exemplos detalhados e raciocínio explícito eleva a quantidade de *tokens* enviados ao modelo. Essa exigência de equilíbrio entre detalhamento e economia de contexto pode ter limitado a profundidade das orientações oferecidas, comprometendo a qualidade da classificação em determinados casos. Assim, os resultados devem ser analisados considerando essas ameaças, que não comprometem o estudo, mas indicam direções claras para melhorias futuras, particularmente no que diz respeito ao tratamento de elementos visuais e à distinção de questões com similaridade estrutural.

8. Conclusão

Este estudo propôs um método para a classificação automática de questões da OBI (modalidade Iniciação) com o uso da técnica de *CoT prompting*, baseada nas classes e tipos estabelecidos por Martins [2011]. Foram criados diversos *prompts*, passando de estratégias *zero-shot* para versões com raciocínio explícito, com o objetivo de analisar como a estrutura do *prompt* afeta a precisão da classificação.

Os experimentos mostraram que a utilização do raciocínio passo a passo da técnica trouxe melhorias significativas no desempenho em comparação com abordagens mais básicas, como demonstrado na Seção 5.1. A taxa de acertos aumentou de forma consistente, o que indica que essa técnica é eficaz para lidar com a estrutura lógica e o significado das perguntas, mesmo quando há mais elementos visuais ou mudanças na estrutura das questões. Além disso, os resultados confirmam que a classificação proposta por Martins(2011, p.23) ainda é adequada para categorizar as questões das provas mais atuais, sem a necessidade de criar novas categorias ou tipos de perguntas.

De forma geral, os resultados obtidos reforçam o potencial de técnicas modernas de *prompting*, em especial o *CoT*, como ferramentas de apoio à análise educacional e à construção de sistemas inteligentes para classificação automática de questões.

Como trabalhos futuros, novas melhorias e atividades que podem ser realizadas, temos por exemplo:

- Modificações para melhorar a precisão do *prompt* final;
- Modificações para permitir a análise de imagens de uma questão;
- Aplicação da atividade em provas de outras olimpíadas de conhecimento.
- Aplicação da atividade em provas posteriores, com o intuito de verificar possíveis novas classes ou tipos.

Referências

Al Faraby, S., Romadhony, A., and Adiwijaya (2024). Analysis of llms for educational question classification and generation. *Computers and Education: Artificial Intelligence*, 7:100298.

IC-UNICAMP (2000). Olimpíada brasileira de informática. `https://olimpiada.ic.unicamp.br/`. Acesso em 13 de maio de 2025.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., and Kasneci, G. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

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

OpenAI (2022). Introducing chatgpt. https://openai.com/index/chatgpt/. Acesso em 28 de maio de 2025.

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

Phoenix, J. and Taylor, M. (2024). *Prompt Engineering for Generative AI*. O'Reilly Media.

Ramlochan, S. (2023). Master prompting concepts: Zero-shot and few-shot prompting. https://promptengineering.org/master-prompting-concepts-chain-of-thought-prompting/. Acesso em 13 de maio de 2025.

Trindade, R. G., da Silveira Schneider, C. A. D., and Charao, A. S. (2017). Logic in a logic way: um aplicativo para exercitar a resolução de problemas de lógica da olimpíada brasileira de informática. In *Workshop sobre Educação em Computação (WEI)*, pages 2237–2246. SBC.

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.

Wolff, T. (2023). How to craft prompts for maximum effectiveness. https://tristwolff.medium.com/from-zero-shot-to-chain-of-thought-prompt-engineering-choosing-the-right-prompt-types-88800f242137. Acesso em 30 de abril de 2025.