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 (FACOM)
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. *This paper focuses on an analysis of the structure of questions in the OBI (Brazilian Informatics Olympiad) initiation category and explores how Chain-of-Thought prompting can be used for question classification, particularly to assess whether the currently available study materials are compatible with the most recent exams.*

Resumo. *Este artigo foca em uma análise da estrutura das questões da OBI na modalidade iniciação e como podemos utilizar Chain-of-thought prompting para a classificação das questões, particularmente para entendermos se os materiais de estudo atuais disponíveis são compatíveis com as provas mais recentes.*

1. Introdução

A Olimpíada Brasileira de Informática (OBI) [IC-UNICAMP 2000] é uma competição organizada nos moldes das outras olimpíadas científicas brasileiras, a OBI é organizada 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). No site da olimpíada existe uma seção de preparação, onde o único material indicado para a preparação da prova é o livro Jogos de Lógica [Martins 2011]. Em seu livro [Martins 2011], são descritas classes e tipos de questões, bem como seus métodos de resolução, contudo o conjunto de questões contempladas, são as questões da OBI 2003 a OBI 2009.

Diante desse fato, esse artigo aborda a utilização da técnica Chain-of-Thought(CoT) prompting para classificação de questões de edições mais recentes da olimpíada e como auxílio para avaliar a adequação das classes e tipos apresentadas no livro Jogos de Lógica [Martins 2011].

2. Fundamento Teórico

Modelos de linguagem de grande escala (LLMs), como o GPT-4, 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, LLMs 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].

2.1. 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.

2.2. One-Shot Prompting

Essa técnica disponibiliza ao modelo um exemplo para guiar a resolução da tarefa antes de apresentar a tarefa em si. 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.

2.3. 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 explicito 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.

3. 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.

3.1. Ordenação

- **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.

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

4. Metodologia

Para assegurar que o prompt escrito fosse eficaz no auxilio da classificação das questões, os seguintes passos foram seguidos:

4.1. Análise manual do dataset

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

4.1.1. Estrutura das questões

Durante a revisão inicial das questões foi observado que ao longo das aplicações da OBI pouco foi alterado em relação ao formato das questões, então podemos identificar uma estrutura básica:

- **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 resolução da questão

- **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.

4.2. Escrita do prompt para a identificação das questões

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. 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.

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 1. Excerto do prompt Zero-Shot (Classe)

4.3. Aplicação dos tipos no prompt inicial

Após a comparação com a análise manual, o prompt zero-shot foi modificado 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.

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 2. Excerto do prompt Zero-Shot (Classe-Tipo)

4.4. Aplicação de exemplos de classe - tipo

A próxima etapa foi a inserção de exemplos de questões classificadas em cada tipo no prompt. 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 auxilio 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.”

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 3. Excerto do prompt One-Shot

4.5. Utilização do chain-of-thought prompting

Por fim, foi escrita a versão do prompt contendo CoT, adicionando a linha de pensamento de classificação dos exemplos de cada tipo. Os dados foram comparados com os outros até então obtidos. A Figura 4 utiliza como base a questão da Figura 3.

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 4. Excerto do prompt CoT

5. Resultados

5.1. Desempenho dos prompts

No total foram classificadas 899 questões, 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

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”.

Outro 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.

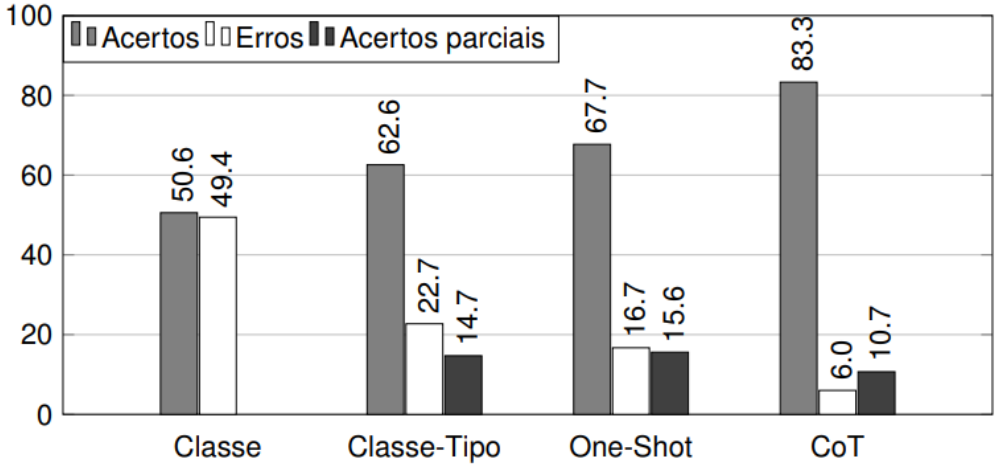

Figura 5. Gráfico de resultado da classificação das provas de 2003 a 2009

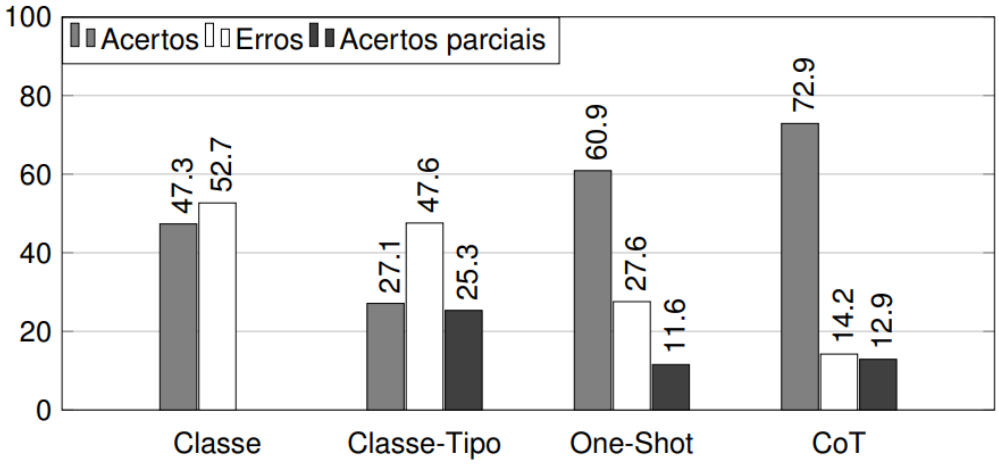

Figura 6. Gráfico de resultado da classificação das provas de 2023 a 2024

5.2. Avaliação das Classes e Tipos

Sobre as questões da prova em si, foi notado um aumento na utilização de recursos visuais em todas as partes que compõem uma questão (contexto, regras e perguntas), apesar desse fato as classes e tipos apresentados no livro [Martins 2011], continuam sendo adequadas para classificação precisa das questões.

6. Conclusão

Este estudo propôs um método para a classificação automática de questões da OBI (modalidade Iniciação) com o uso da técnica de Chain-of-Thought 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 ainda é adequada para categorizar as questões das provas mais atuais, sem a necessidade de criar novas categorias ou tipos de perguntas por enquanto.

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 podem ser realizadas, 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 posteriores, com o intuito de verificar possíveis novas classes ou tipos.

Referências

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

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.

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.

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.