Estudo de caso: uso do ChatGPT para resolução de problemas de programação

Débora Souza[1], Rohit Gheyi[1]

[1]Universidade Federal de Campina Grande (UFCG)

debora.souza@ccc.ufcg.edu.br, rohit@dsc.ufcg.edu.br

Abstract. *The generation of programs from natural language aims to transform sentences into programming code. ChatGPT, based on the GPT-3 model, is a chatbot capable of generating text and code like a human. To assess its reliability in the programming context, 100 programming problems were randomly selected, and distributed among Easy, Intermediate, and Hard levels of complexity from popular platforms for the model to solve. Over 3 attempts, ChatGPT correctly solved 71 problems, with 50 from the LeetCode platform and 21 from the BeeCrowd platform. The remaining 29 unsolved problems were mostly distributed among Intermediate and Hard levels of complexity. In this context, the study demonstrates that the use of ChatGPT can be beneficial for programmers as the model can provide examples of well-crafted code and offer solutions to their challenges. However, it also highlights that ChatGPT does not achieve 100% accuracy; therefore, programmers should not consider it a complete replacement.*

Resumo. *A geração de programas a partir de linguagem natural visa transformar frases em código de programação. O ChatGPT, baseado no modelo GPT-3, é um chatbot capaz de gerar texto e código como humano. Para avaliar sua corretude no contexto de programação, foram selecionados aleatoriamente 100 problemas de programação, distribuídos entre os graus de complexidade Fácil, Intermediário e Difícil, de plataformas populares para serem resolvidas pelo modelo. Ao longo de 3 tentativas, o ChatGPT acertou 71 problemas, sendo 50 da plataforma LeetCode e 21 da plataforma BeeCrowd, os 29 problemas não solucionados estão majoritariamente distribuídos entre os graus de complexidade Intermediário e Difícil. Nesse contexto, o estudo demonstra que a utilização do ChatGPT pode ser benéfica para os programadores, pois o modelo é capaz de fornecer exemplos de código com soluções para seus desafios. No entanto, também evidencia que o ChatGPT não atinge uma precisão de 100%, portanto, os programadores não devem considerá-lo como um substituto completo.*

1. Introdução

A geração de programas a partir de linguagem natural utiliza algoritmos de processamento de linguagem natural e aprendizado de máquina para converter frases ou comandos em código de programação. Essa tecnologia visa tornar a codificação mais acessível e intuitiva para pessoas sem conhecimento de programação, aumentando a eficiência e reduzindo erros [1].

Um exemplo de ferramenta de geração de código é o GitHub Copilot [2], lançada em 2021, que visa ajudar os desenvolvedores a escrever código mais rapidamente e com mais eficiência. Outra ferramenta é o ChatGPT [3], um chatbot de propósito geral baseado no modelo de linguagem GPT-3 desenvolvido pela empresa OpenIA. Foi desenvolvido para gerar texto e código como humano, e treinado em uma forma de conversação usando aprendizagem por reforço com feedback humano. O

ChatGPT já se mostrou capaz de realizar diversas atividades relacionadas à escrita [4]. Nesse cenário, surgiram dúvidas sobre a substituição parcial ou total da mão de obra humana em diversos setores, bem como questionamentos acerca da corretude das respostas fornecidas pelo ChatGPT.

Este trabalho visa avaliar a corretude do ChatGPT na resolução de problemas de programação, assim como as vantagens e limitações do seu uso. Nós avaliamos o desempenho do ChatGPT 3.5 ao resolver problemas de programação das plataformas LeetCode e BeeCrowd, que são populares para prática e aprimoramento de habilidades de programação. O LeetCode é uma plataforma amplamente utilizada para questões de programação, englobando problemas utilizados em processos seletivos de empresas de tecnologia [5]. O portal LeetCode está atualmente entre as melhores plataformas para avaliar conhecimento sobre programação, com questões que abordam os mais variados conceitos de computação. Já o BeeCrowd é uma plataforma focada em competições de programação de alta dificuldade, especialmente em formato de maratonas em equipe [6]. A plataforma disponibiliza desafios de alta dificuldade e também inclui problemas das edições anteriores da Olimpíada Brasileira de Informática e da Maratona de Programação.

Para o nosso estudo, foram selecionados aleatoriamente 50 problemas de cada plataforma distribuídos entre os graus de complexidade Fácil, Intermediário e Difícil, podendo ou não conter imagem ilustrativa, para serem submetidos ao ChatGPT. Ao longo de 3 tentativas, o ChatGPT acertou 71 problemas, sendo 50 da plataforma LeetCode e 21 da plataforma BeeCrowd. Dos 100 problemas selecionados, 21 deles continham imagens ilustrativas. O ChatGPT resolveu corretamente 17 deles, mostrando que o modelo lida bem com o texto do enunciado quando este é suficiente para descrever o problema. O objetivo do estudo é avaliar a corretude do ChatGPT na geração automática de código para desenvolvedores.

Esse artigo está organizado da seguinte forma: a Seção 2 trata da metodologia, a Seção 3 apresenta os resultados e discussões, a Seção 4 trata dos trabalhos relacionados, e por fim a Seção 5 apresenta as conclusões e trabalhos futuros.

2. Metodologia

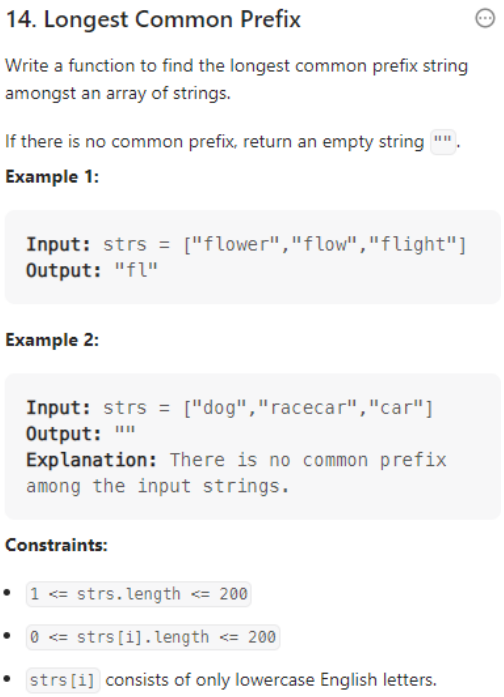
GQM. O objetivo deste trabalho é analisar as respostas geradas pelo ChatGPT para resolução de problemas de programação, com o propósito de levantar implicações do seu uso no dia a dia de programadores, com respeito a corretude das respostas geradas, do ponto de vista de pesquisadores, no contexto de geração de código automático. Para isso responderemos a seguinte questão de pesquisa (QP):

- QP_1. Até que ponto o ChatGPT é capaz de resolver problemas de plataformas de programação?
    - Serão contadas as respostas corretas e incorretas fornecidas pelas plataformas.

Para avaliar se o ChatGPT responde corretamente problemas de programação, 100 problemas das plataformas LeetCode e BeeCrowd foram aleatoriamente selecionados e manualmente submetidos ao ChatGPT. Os problemas estão distribuídos entre os graus de complexidade Fácil, Intermediário e Difícil e são relacionados a

diversas áreas da computação, como arrays, estruturas de dados, algoritmos e grafos, entre outras.

As ferramentas LeetCode e BeeCrowd foram escolhidas por serem repositórios populares de questões que são usadas em entrevistas. Sendo assim, elas podem ajudar a simular problemas reais que as empresas possam ter no dia a dia. A Figura 1 apresenta um exemplo de problema selecionado, que contém a descrição, exemplos com valores de entrada e saída e as restrições a respeito das variáveis de entrada.



14. Longest Common Prefix

Write a function to find the longest common prefix string amongst an array of strings.

If there is no common prefix, return an empty string `""`.

Example 1:

```
Input: strs = ["flower","flow","flight"]
Output: "fl"
```

Example 2:

```
Input: strs = ["dog","racecar","car"]
Output: ""
Explanation: There is no common prefix
among the input strings.
```

Constraints:

- `1 <= strs.length <= 200`
- `0 <= strs[i].length <= 200`
- `strs[i]` consists of only lowercase English letters.

Figura 1. Exemplo de problema selecionado da plataforma LeetCode.

Sendo assim, para avaliação do problema pelo ChatGPT, todo o texto (descrição, exemplos e restrições) do problema foi manualmente copiado da plataforma e colado no *prompt* para o chat responder. Este, por sua vez, retorna uma solução que é submetida na plataforma, e avaliada como aceita ou rejeitada de acordo com os casos de testes disponíveis na plataforma.

Apesar do exemplo na Figura 1 não conter uma imagem explicativa, é comum que os enunciados contenham imagens para elucidar o problema, no entanto, este artefato não foi passado para ser avaliado junto a questão para o ChatGPT 3.5, já que não possui suporte para interpretação de imagens.

A plataforma LeetCode divide seus problemas em níveis de dificuldade, sendo eles Fácil, Intermediário e Difícil. Sendo assim, 50 problemas foram selecionados onde são 20 do nível fácil, 15 do nível intermediário e 15 do nível difícil. Já a plataforma BeeCrowd divide seus problemas por nível de dificuldade de 1 a 10, sendo assim, 5 questões de cada nível foram selecionadas, totalizando 50 problemas. Para facilitar a comparação com os problemas do LeetCode, as questões foram redistribuídas em três níveis semânticos, sendo eles:

- Fácil: questões do nível 1 ao 4;

- Intermediário: questöes do nível 5 ao 7;
- Difícil: questões do nível 8 ao 10.

É importante ressaltar que no BeeCrowd o nível de dificuldade é estimado usando uma variação do ELO Rating System [7] que usa o número de vezes que o problema foi "derrotado" (ou seja, quantos usuários resolveram o problema) para determinar sua dificuldade. Nesse sistema, problemas que são resolvidos por vários usuários com poucas tentativas são classificados como de baixa dificuldade, enquanto problemas com mais tentativas, mas menos resolvidos, são classificados como de alta dificuldade. A dificuldade dos problemas é reajustada semanalmente, portanto, sua pontuação total pode variar [8]. Os problemas utilizados para essa análise foram selecionados em 01/04/2023.

Dado que o ChatGPT pode fornecer respostas variadas para uma única pergunta, o ChatGPT recebeu três oportunidades para resolver cada problema. Isso significa que o mesmo texto foi submetido no máximo três vezes para avaliação. A decisão de permitir três submissões é completamente experimental, sendo que em cada submissão, apenas o enunciado da questão é encaminhado para o chat. Os *prompts* foram os mesmos nas três tentativas. Se após a terceira submissão a resposta ainda for incorreta, então o problema é identificado como não solucionado. Cada problema foi submetido individualmente ao modelo em um *prompt* exclusivo, o *prompt* foi reutilizado apenas quando houve a necessidade da segunda e terceira submissão. Nenhuma linguagem de programação específica foi designada no *prompt* para o ChatGPT, no entanto, todas as perguntas foram respondidas utilizando Python.

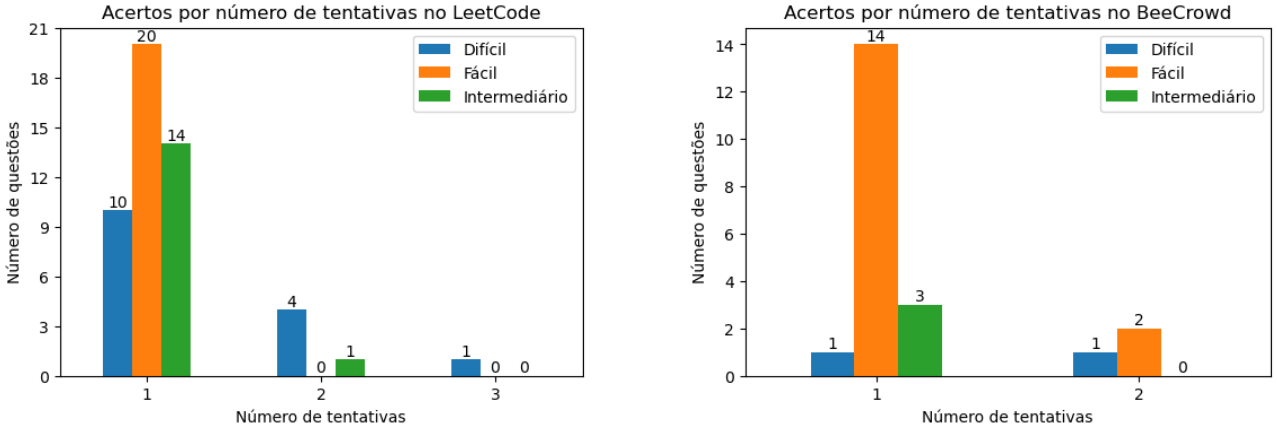
Para cada problema avaliado, foi documentado em uma planilha para posterior análise seu enunciado, nível de dificuldade, código de solução proposto pelo ChatGPT, o *status* de corretude da submissão, o erro lançado pela plataforma no caso da resposta ser incorreta, o(s) conteúdo(s) abordado(s) na questão, etc. Essa planilha está disponível online [15]. A versão do ChatGPT usada foi treinada com dados até setembro de 2021. Não foi utilizada a versão 4 do ChatGPT pois enquanto esta pesquisa estava sendo feita a plataforma não estava disponível gratuitamente.

3. Resultados e Discussão

Na plataforma LeetCode, o ChatGPT obteve um ótimo desempenho na resolução de problemas. Foram submetidos 50 problemas, onde são 20 do nível fácil, 15 do nível intermediário e 15 do nível difícil. Após a submissão e avaliação das respostas coletadas, todas as 50 questões foram respondidas corretamente após no máximo 3 tentativas.

Dos 50 problemas submetidos ao LeetCode, 16 deles (32%) continham imagens ilustrativas, sendo 6 do nível Fácil, 5 do nível Intermediário e 5 do nível Difícil, no entanto, o ChatGPT 3.5 não as processou. Apesar disso, 14 desses problemas foram resolvidos corretamente na primeira tentativa, indicando que o modelo lida bem com o texto do enunciado quando este é suficiente para entender o problema. Os 2 problemas remanescentes exigiram uma segunda submissão e foram solucionados, nenhum problema com imagem exigiu uma terceira submissão.

De acordo com a Figura 2A apresentada a seguir, dos 50 problemas selecionados, aproximadamente 88% deles foram resolvidos corretamente na primeira submissão ao ChatGPT. Todos os problemas de nível fácil foram solucionados na primeira tentativa, bem como cerca de 93% dos problemas de nível intermediário e aproximadamente 67% dos problemas de nível difícil. Os problemas que não foram resolvidos na primeira tentativa foram submetidos novamente ao ChatGPT. Na segunda rodada de submissões, o ChatGPT conseguiu solucionar 100% dos problemas restantes do nível intermediário, bem como 80% dos problemas restantes do nível difícil. Por fim, na terceira rodada de submissões, o ChatGPT conseguiu resolver 100% dos problemas remanescentes do nível difícil.



(A) **(B)**

Figura 2. A: Número de acertos por número de tentativas da plataforma LeetCode. B: Número de acertos por número de tentativas da plataforma BeeCrowd.

Na plataforma BeeCrowd, o ChatGPT obteve um desempenho moderado. Assim como na plataforma LeetCode, foram submetidos 50 problemas, sendo 20 do nível fácil, 15 do nível intermediário e 15 do nível difícil. Após a submissão e avaliação das respostas coletadas, obtemos que das 50 questões respondidas pelo ChatGPT, apenas 21 delas, cerca de 42%, foram respondidas corretamente, como mostra a Figura 3 a seguir.

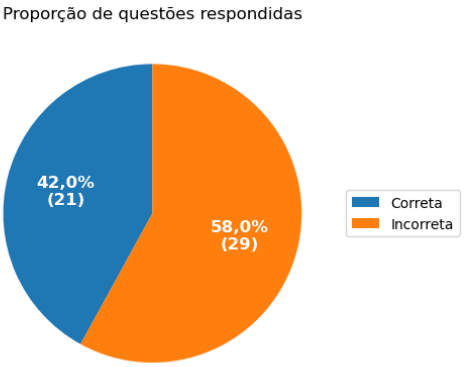


Figura 3. Proporção de problemas respondidos na plataforma BeeCrowd.

Dos 50 problemas submetidos à plataforma BeeCrowd, apenas 5 deles (10%) possuíam imagens ilustrativas no enunciado. Desses, 4 problemas eram de nível fácil,

enquanto 1 era de nível difícil. Ao submeter os problemas para o ChatGPT, as imagens foram desconsideradas, e apenas o texto do enunciado foi utilizado. Entre os problemas de nível fácil, 3 foram solucionados corretamente, independentemente das imagens, enquanto 2 problemas não foram resolvidos corretamente, sendo 1 de nível fácil e 1 de nível difícil. Devido à escolha aleatória dos problemas e à baixa quantidade de problemas com imagens (apenas 5), não dispomos de dados suficientes para determinar se as imagens afetaram ou não o desempenho do ChatGPT na resolução dos problemas.

Ao avaliar a distribuição do número de questões corretas e incorretas por grau de dificuldade, é possível verificar como foram os resultados para problemas de nível fácil, intermediário e difícil. Conforme ilustrado na Figura 4 abaixo, dos 20 problemas de nível fácil apresentados, aproximadamente 80% deles foram respondidos corretamente, totalizando 16 acertos. Em contrapartida, as questões de níveis intermediário e difícil tiveram um desempenho inferior, com apenas 3 e 2 problemas corretamente respondidos, o que equivale a cerca de 20% e 13%, respectivamente.

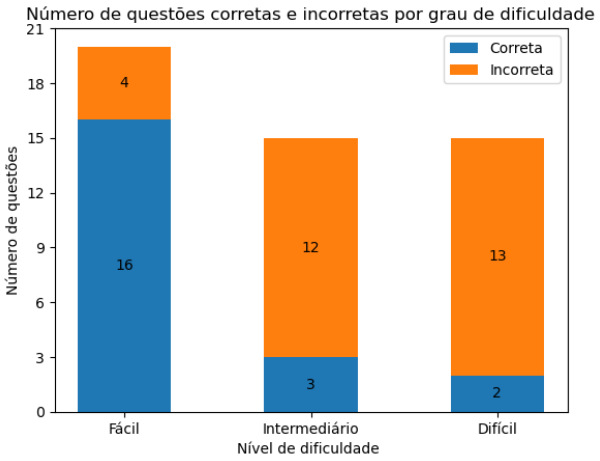


Figura 4. Número de questões corretas e incorretas por grau de dificuldade na plataforma BeeCrowd.

Uma vez que cada problema foi submetido no máximo três vezes, é possível analisar a proporção de acertos em relação ao número de tentativas. De acordo com a Figura 2B apresentada, apenas 36% dos problemas foram resolvidos corretamente na primeira tentativa. Isso significa que 70% dos problemas de nível fácil foram solucionados corretamente na primeira rodada de submissões, bem como 20% dos problemas de nível intermediário e apenas 7% dos problemas de nível difícil. Os problemas que não foram solucionados na primeira tentativa foram submetidos novamente ao ChatGPT. Na segunda rodada de submissões, 33% dos problemas restantes de nível fácil foram solucionados, bem como 7% dos problemas remanescentes de nível difícil. Entretanto, nenhum problema de nível intermediário foi solucionado nesta rodada. Por fim, na terceira rodada de submissões, nenhum problema foi solucionado.

Os problemas selecionados da plataforma BeeCrowd originam-se de diversas fontes. Como as questões selecionadas para este estudo foram escolhidas de forma aleatória, suas origens também são variadas. Portanto, não foi possível estabelecer uma correlação entre problemas não resolvidos e uma fonte específica.

A plataforma BeeCrowd fornece uma saída contendo o erro associado para cada problema que não é solucionado. Sendo assim, é possível analisar os tipos de erros gerados e sua proporção de ocorrência. Como mostra a Tabela 1 a seguir, quatro erros ao total foram lançados, os principais erros produzidos foram *Time Limit Exceeded* e *Wrong Answer* (100%). *Time Limit Exceeded* é um erro lançado quando a solução enviada leva mais tempo do que o permitido para executar todos os testes de avaliação [9]. Já o *Wrong Answer* (100%) é lançado quando a solução não apresenta o resultado esperado para 100% dos casos de teste.

Além disso, há também o erro *Runtime Error*, lançado em menor quantidade, que diz respeito a definir um vetor ou array com capacidade inferior à necessária para o problema, ou ao tentar acessar uma posição inválida na memória [9]. Já *Memory Limit Exceeded*, erro lançado apenas uma vez, é gerado quando o código tenta alocar mais memória do que o máximo permitido para o problema. Isso pode ocorrer porque está sendo utilizado um vetor ou uma estrutura de dados muito grande [9].

Erro gerado	Contagem do erro
Time limit exceeded	12
Wrong answer (100%)	10
Runtime error	6
Memory limit exceeded	1

Tabela 1. Tipos de erros gerados pelo BeeCrowd e suas frequências de ocorrência.

Sendo assim, é possível afirmar que, em sua maioria, as respostas incorretas geradas pelo ChatGPT são rejeitadas por não terem sido suficientemente otimizadas para serem aprovadas ou por não terem passado em nenhum caso de teste.

As plataformas LeetCode e BeeCrowd fornecem informações sobre os tópicos abordados em cada questão, permitindo assim a avaliação de quais conteúdos foram abordados. As Tabelas 2A e 2B a seguir exibem os principais tópicos tratados nas questões discutidas neste estudo.

Tópico	Número de Perguntas
array	17
string	17
math	10
hash table	9
two pointers	9
linked list	6
dynamic programming	6
recursion	6
depth-first search	5
tree	5

(A)

Tópico	Número de Perguntas
ad-hoc	11
beginner	9
data structures and libraries	8
math	7
paradigms	6
graph	5
strings	4

(B)

Tabela 2. A: Principais tópicos abordados nas questões da plataforma LeetCode. B: Principais tópicos abordados nas questões da plataforma BeeCrowd.

Os problemas na plataforma LeetCode estão associados a um ou mais tópicos, conforme demonstrado na Tabela 2A, que apresenta os 10 tópicos mais frequentemente abordados. Isso explica por que o número de tópicos contados excede o número de problemas selecionados (50 problemas). É notável que os tópicos abordados são amplamente reconhecidos no mundo da programação, incluindo *Array, String e Math*. Alguns problemas também fazem uso de conteúdos mais complexos, como programação dinâmica e recursão. Ainda assim, o ChatGPT acertou 100% das questões submetidas.

Já na plataforma BeeCrowd, cada problema está vinculado a apenas um tópico, como evidenciado na Tabela 2B, que exibe todos os tópicos abordados nas questões selecionadas para este estudo. Os conteúdos abordados nessa plataforma também são comuns ao mundo da programação, como *Data Structures and Libraries, Math e Paradigms*. Os tópicos mais abordados são *Ad-Hoc* e *Beginner*, categorias criadas pela plataforma para direcionar problemas que não se encaixam nas demais categorias, e são básicos o suficiente iniciantes na programação [14], respectivamente. Ao examinarmos os tópicos tratados, fica evidente que as questões da plataforma BeeCrowd não abordam conteúdos significativamente distintos ou mais complexos em comparação com as questões da plataforma LeetCode. Isso sugere que a falta de respostas adequadas do ChatGPT não se deve ao conteúdo das questões, mas sim à formulação das perguntas.

Ameaças à Validade. A respeito das avaliações das submissões, assumimos que as respostas das plataformas são corretas. Essa abordagem se baseia na ampla utilização dessas plataformas por diversos desenvolvedores. Qualquer falha ou erro na correção das questões provavelmente seria identificado e relatado pela comunidade, o que permitiria a correção oportuna desses problemas. Além disso, o ChatGPT realiza buscas regulares na internet para manter suas respostas atualizadas, por isso, as soluções apresentadas podem ser atualizadas, permitindo ao ChatGPT corrigir respostas que anteriormente podem ter sido imprecisas.

4. Trabalhos Relacionados

Kabir et al. [18] analisaram as respostas do ChatGPT a 517 perguntas do Stack Overflow para avaliar a exatidão, consistência, abrangência e concisão das respostas do ChatGPT 3.5. Esta pesquisa mostrou que 52% dessas respostas contêm imprecisões e 77% são verborrágicas. Khoury et al. [19] realizaram um estudo para avaliar a segurança do código gerado pelo ChatGPT 3.5. Eles solicitaram a geração de 21 programas e scripts e descobriram que, em muitos casos, o código não atendia aos padrões mínimos de segurança. O ChatGPT também reconheceu que o código produzido não era seguro quando questionado. Em nosso trabalho, nós avaliamos a corretude do ChatGPT 3.5 analisando a corretude das soluções para 100 problemas das plataformas LeetCode e BeeCrowd.

5. Conclusão

Neste trabalho, conforme evidenciado, o ChatGPT tem um desempenho satisfatório, especialmente em problemas de complexidade média e baixa, no entanto, apresenta certa dificuldade em problemas considerados difíceis. O desempenho insatisfatório do ChatGPT na plataforma BeeCrowd pode ser atribuído à grande variedade de

especificações. A maioria dos problemas nessa plataforma são de competições e apresentam enunciados que usam histórias fictícias e lúdicas, envolvendo personagens e exemplos, muitas vezes de forma intencionalmente vaga para aumentar a complexidade do texto, e com isso dificultar o entendimento pelos competidores. Por outro lado, na plataforma LeetCode, os enunciados são breves e objetivos com o objetivo de avaliar um determinado conteúdo.

Implicações. O uso do ChatGPT pode ser vantajoso para programadores, uma vez que o modelo é capaz de fornecer exemplos de códigos com soluções para seus problemas. Mas como foi evidenciado, o ChatGPT não possui 100% de corretude, por isso, os programadores não devem enxergá-lo como um substituto. Alguns especialistas comentam que o ChatGPT não tem a capacidade de desenvolver um código complexo, ou garantir que não haverá bugs e será totalmente seguro, de fácil manutenção e bem documentado, mas que a ferramenta deve servir como um complemento ao setor de tecnologia [10].

Já para estudantes, há a vantagem de ter a sua disposição um auxiliar para explicar conteúdos e esclarecer dúvidas, mas também há desvantagens, uma vez que o ChatGPT pode responder erroneamente e o estudante não estará capacitado para detectar o erro. Em resumo, o ChatGPT pode ser muito útil para agilizar o desenvolvimento de software. Mas, é importante que o programador seja responsável, atencioso e crítico o suficiente para avaliar a resposta gerada e não utilizá-la sem fazer os devidos testes.

Trabalhos Futuros. Como sugestão para trabalhos futuros, é possível explorar o potencial da versão ChatGPT 4 (indisponível gratuitamente no momento desse estudo), que pode oferecer um melhor desempenho na resolução de problemas mais complexos, graças ao seu conhecimento geral mais amplo e habilidades de resolução de problemas [11], além de aceitar o input de imagens. Outra possibilidade seria realizar o estudo considerando outras plataformas de problemas, como HackerRank e Codeforces, com o intuito de identificar um padrão de estrutura de problema no qual o ChatGPT possa ter um melhor desempenho. Uma abordagem similar pode ser realizada para avaliar a corretude na descoberta de bugs [16] e identificação de problemas na compreensão do código [17], como também avaliar modelos para ver até que ponto o ChatGPT ajuda também na síntese e evolução de especificações [12].

Referências

[1] ChatGPT. Disponível em: https://openai.com/blog/chatgpt. Acesso em: Setembro 2023.

[2] Getting started with GitHub Copilot. Disponível em: https://docs.github.com/en/copilot/getting-started-with-github-copilot. Acesso em: Setembro 2023.

[3] ChatGPT. Disponível em: https://openai.com/chatgpt. Acesso em: Setembro 2023.

[4] Recruitment team unwittingly recommends ChatGPT for job interview. Disponível em:

https://news.sky.com/story/recruitment-team-unwittingly-recommends-ChatGPT-for-job-interview-12788770. Acesso em: Setembro 2023.

[5] Plataforma LeetCode. Disponível em: https://leetcode.com/. Acesso em: Setembro 2023.

[6] Plataforma BeeCrowd. Disponível em: https://BeeCrowd.com.br/. Acesso em: Setembro 2023.

[7] ELO Rating System. Disponível em: https://medium.com/purple-theory/what-is-elo-rating-c4eb7a9061e0. Acesso em: Setembro 2023.

[8] FAQS Judge - BeeCrowd. Disponível em: https://www.beecrowd.com.br/judge/en/faqs. Acesso em: Setembro 2023.

[9] FAQS Answer - BeeCrowd. Disponível em: https://www.beecrowd.com.br/judge/en/answers. Acesso em: Setembro 2023.

[10] Concorrente ou aliado? Especialista defende que ChatGPT pode alavancar a carreira de programador. Disponível em: https://exame.com/carreira/concorrente-ou-aliado-especialista-defende-que-chatgpt-pode-alavancar-a-carreira-de-programador/. Acesso em: Setembro 2023.

[11] GPT-4. Disponível em: https://openai.com/product/gpt-4. Acesso em: Setembro 2023.

[12] Rohit Gheyi, Paulo Borba. Refactoring Alloy Specifications. In Cavalcanti, A., Machado, P., eds.: Electronic Notes in Theoretical Computer Science, Proceedings of the Brazilian Workshop on Formal Methods. Volume 95. Elsevier (2004) 227–243.

[14] Categorias dos problemas da plataforma BeeCrowd. Disponível em: https://www.beecrowd.com.br/judge/en/categories. Acesso em: Setembro 2023.

[15] Débora Souza, Rohit Gheyi. Questões submetidas ao ChatGPT. Disponível em: https://docs.google.com/spreadsheets/d/1jMuFXYphU6lDffXJ0mfUHsNsMz_yVhqcbzlF1AqTqA0/edit?usp=sharing. Acesso em: Setembro 2023.

[16] F. Medeiros et al.. An empirical study on configuration-related issues: investigating undeclared and unused identifiers. GPCE 2015, pages 35–44, 2015.

[17] F. Medeiros et al.. Investigating misunderstanding code patterns in C open-source software projects. Empirical Software Engineering, 24:1693–1726, 2019.

[18] Samia Kabir, David N. Udo-Imeh, Bonan Kou, Tianyi Zhang. Who Answers It Better? An In-Depth Analysis of ChatGPT and Stack Overflow Answers to Software Engineering Questions. Disponível em: https://arxiv.org/abs/2308.02312. Acesso em: Setembro 2023.

[19] Raphaël Khoury, Anderson Avila, Jacob Brunelle, Baba Camara. How Secure is Code Generated by ChatGPT?. Disponível em: https://arxiv.org/abs/2304.09655. Acesso em: Setembro 2023.