Misinformation and Disinformation in ChatGPT

FIT1055 | 12 May 2023

Chat Overflow

Brandon Lau | 32184654

Chen Xi Diong | 32722656

Yi Bin Sia | 33363129

Lana Hussayni | 33267871

Shunnosuke Takei | 32843909

Hong Bo Kang | 32684673

1. Background

A product of generative AI, OpenAI's GPT-3 language model, also known as ChatGPT, is capable of generating human-like responses to prompts and questions. It was developed and launched as a prototype in November 2022. ChatGPT's user-friendly design and intuitive interface make the power of AI accessible to anyone, regardless of their technical expertise. Since the language model of ChatGPT is built using techniques of Natural Language Processing and Natural Language Generation, the AI can understand and provide human-like text responses. It can also be prompted to write in a specific style, according to research by Dehouche (2021).

One important thing to note is that the data used to train ChatGPT is obtained by crawling the web and scraping text from various websites, such as news sites, social media, and other online sources. This data may be false or inaccurate as it is only sanitised based on its irrelevance and quality, and the credibility of the sources is not checked. Additionally, most users do not validate the responses provided by ChatGPT and blindly believe what they are told is real.

The ethical issues pertaining to ChatGPT are the usage of the chatbot to create misinformation and disinformation, which is not a small issue given how big the user base has grown since its release. Misinformation refers to false or inaccurate information that is shared unintentionally, often due to a lack of knowledge or incorrect assumptions. Disinformation refers to false or misleading information that is deliberately spread with the intention of deceiving people or manipulating public opinion. With the ability to generate convincing responses, ChatGPT is a powerful tool to provide false information, both intentionally and unintentionally. This feature of ChatGPT can be leveraged by different entities to conduct illegal activities, causing harm to various stakeholders of the tool.

In this report, we will identify the stakeholders who are negatively affected by this ethical issue and the ACM codes that are violated. Using the ethical theories as a guideline, we will also propose a solution that serves to alleviate the severity of the situation, all while adhering to the Ethical Reasoning Framework (ERF).

2. Methodology

2.1 ERF

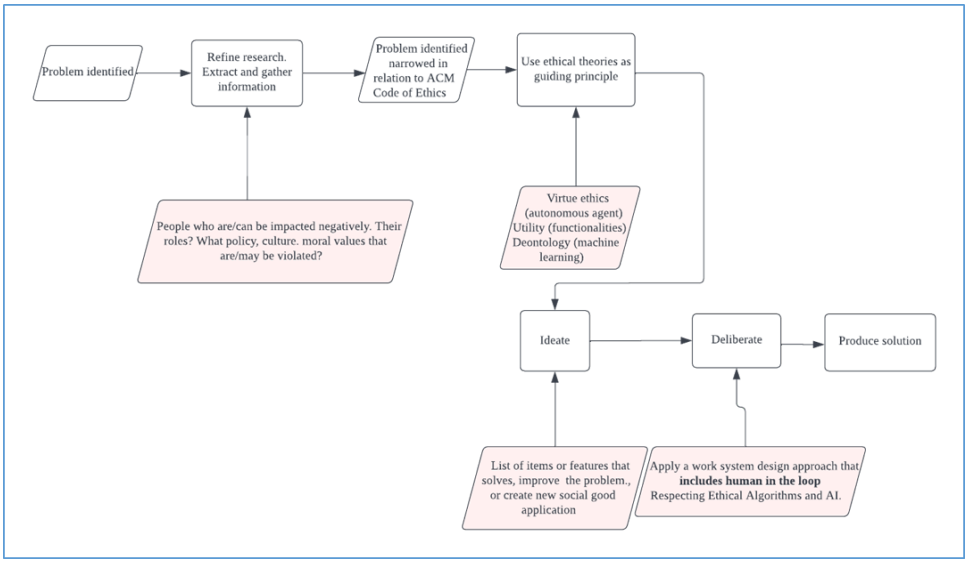

Figure: Flowchart of the Ethical Reasoning Framework (Nailah 2022)

Ethical Reasoning Framework (ERF) is defined as a foundation that consists of a set of methods which one uses to decide what to do or how to develop an ethical algorithm and/or AI application (Nailah, 2022). The framework involves 7 steps, including:

1. Problem identified: In this step, we identified the ethical problem that needs to be resolved. From our selection of topics of Assignment 1, we narrowed down our topic to be used for this assignment through a discussion.
2. Refine research, extract, and gather information: In this step, we listed down potential factors leading to such an ethical issue, and conducted research on the topic. We manage to identify the stakeholders that are negatively impacted by our problem statement, as well as possible ways to mitigate these issues.
3. Problem identified, narrowed in relation to ACM Code of Ethics: Referring to the ACM Code of Ethics, we identified possible codes that are violated by our problem statement and selected three main codes to further research upon.
4. Use ethical theories as guiding principles: In this step, we use ethical theories as guiding principles to evaluate potential courses of action. This involves using frameworks such as Utilitarianism, Deontology, and Virtue Ethics to spark inspiration for potential solutions. (outcome)
5. Ideate: In this step, We generate a range of possible solutions or options for resolving the ethical problem. (outcome)
6. Deliberate: After generating possible solutions, we deliberate and evaluate each option based on the ethical principles and values identified earlier in the process. (outcome)
7. Problem solution: Finally, we choose the best course of action based on the ethical evaluation and implement the chosen solution. (outcome)

2.1.1 Results

2.1.1.1 Problem Identified

There are four main topics that our group members chose for Assignment 1, namely “Misinformation and Disinformation in ChatGPT”, “Privacy Issues in ChatGPT”, “Misinformation and Disinformation in DeepFake”, and “Privacy Issues in DeepFake”. After a majority vote, we decided to go with the topic of “Misinformation and Disinformation in ChatGPT”. Following the steps of an Intelligent Research Cycle, we laid out the direction that we want to research in, such as the contributing factors to the problem, as well as the ACM Codes violated. This gives us a clear indication of a common goal to work towards as a team.

2.1.1.2 Refine Research, Extract and gather information

We determined that some possible factors of the current issue are ChatGPT has no self-validation algorithms, ChatGPT is trained on large datasets that are prone to inaccuracies and biases, as well as ChatGPT is prone to policy-bypassing prompt injection. We used the Teaming method to carry out research in pairs, and presented the results in our meeting. The research primarily aimed to understand how and why these stakeholders were affected, and resulted in the discovery of potential solutions.

The stakeholders impacted by the issue include community users, the general public, policy-makers, celebrities, programmers and IT developers, government entities, and OpenAI (the company behind ChatGPT's development). For example, OpenAI, being the developers behind the entirety of ChatGPT, it is in their best interests to develop a reliable language model. This issue can negatively impact OpenAI in multiple ways, including but not limited to: loss of partnerships/sponsorships; lawsuits from victims affected by the misuse of ChatGPT; competitors bypassing chatbot policies to gain unauthorised insights on its model. High profile figures such as celebrities and government officials are also negatively affected, due to ChatGPT’s capability in generating fake news in a convincing manner, which oftentimes is a combination of both factual and non-factual information, making it hard to distinguish for the public eye. This can result in defamation and can be the spark of unnecessary conflict.

As a general summary of potential solutions, we think that human control should be introduced to the cycle of ChatGPT usage as much as possible. This is to ensure that ChatGPT operates whilst adhering to the Ethical Algorithms and AI Principles as much as possible.

2.1.1.3 Problem narrowed in relation to ACM Code of Ethics

Referencing from the ACM Code of Ethics and Professional Conduct (Association for Computing Machinery 2018), the main three ACM codes that we believe were violated are:

1.2 Avoid Harm

“Examples of harm include unjustified physical or mental injury, unjustified destruction or disclosure of information, and unjustified damage to property, reputation, and the environment.”

> ChatGPT has been shown to offend this code on multiple different fronts. During our discussion, we have identified that it is able to cause harm towards systems, applications and devices as users are able to prompt ChatGPT to write malware with little to no consequences. ChatGPT is also capable of writing convincing thought pieces that slander public figures. Oftentimes by generating fake news reports and mimicking the target figure’s speech patterns to form a sexist or racist remark. It also causes harm towards individuals as it is capable of writing phishing emails, fooling others into thinking the information presented to them is genuine and approved by the appropriate authorities.

1.4 Be fair and take action not to discriminate

“The values of equality, tolerance, respect for others, and justice govern this principle. Fairness requires that even careful decision processes provide some avenue for redress of grievances.”

> It has already been proven that ChatGPT is heavily influenced by rather controversial ideals when it comes to answering certain questions. It can and often will provide discriminatory answers even when the topic does not seem to have any relation to race, gender etc. An example of this would be ChatGPT generating code that filters for employees that are white and male.

2.5 Give comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks.

“Computing professionals are in a position of trust, and therefore have a special responsibility to provide objective, credible evaluations and testimony to employers, employees, clients, users, and the public.”

> OpenAI demonstrates that they do not explicitly list the disadvantages of ChatGPT to the general public, OpenAI has only stated various policies such as private policy, usage policy, platform policy etc. These policies can be easily bypassed by users, and cause inappropriate responses such as harmful/discriminate content. In general, OpenAI only states what policies it has developed to prevent violations, but does not clearly inform users of the risks they will encounter during actual use.

2.1.1.4 Use ethical theories as guiding principle

Ethical theories, namely Virtue Ethics, Utilitarianism and Deontology are used to formulate guiding principles for us to understand the problem statement better, as well as get inspiration from the said principles to ideate possible solutions later on.

Virtue Ethics:

1. (Honesty) Generative AI models should only be trained using true data, and remind users to fact-check in the case where it is unsure of the credibility of its response.
2. (Honesty and Responsibility) Open AI has the responsibility to show the right information. It shouldn't give information that is not updated
3. Honesty: OpenAI should maintain a level of transparency in its development process. This includes the sources of their training and testing datasets, fixes, etc.
4. (Security) ChatGPT should prioritize the security of user data and take appropriate measures to protect against unauthorized access or data breaches. This includes implementing encryption and other security measures to safeguard user data.

Utilitarianism:

1. When queried for illegal responses such as money laundering or computer hacking, request for user verification and send to authorities for review.
2. Making programs online (i.e. synchronous with real-time) instead of having stagnant data. Outdated information mostly equals false information.
3. Flag responses that contain claims that have been debunked by fact-checking organizations or have been identified as false by reliable sources of information.

Deontology:

1. Generative AI models should refuse to give responses to prompts for illegal activities, no matter what the purpose is.
2. Generative AI can provide alternative responses whenever possible
3. Generative AI shouldn’t influence its users to perform morally wrong actions.
4. Generative AI shouldn't compile false responses to mislead users.
5. Generative AI should prioritize transparency and interpretability, which provide a clear and transparent communication to users about how the AI model works.
6. Generative AI should not lie, and if it detects a question that has not been learned, it should honestly tell the users that it cannot provide an accurate answer.

2.1.1.5 Ideate

Idea	Description	Aims to Resolve
1. Forum website	A community platform for ChatGPT users and developers alike to communicate. Users can raise problems found with the product and raise it for other users or developers to resolve the question. It can also serve as a forum to	● Users are able to flag potential exploits that leverage misinformation and disinformation by using ChatBot. ● Developers are able to learn of policy breaches and

	discuss issues.	vulnerabilities present in the model efficiently.
2. Online AI engine	An AI engine that is connected to the network to provide a more accurate and correct solution for the users.	• To produce a more reliable solution for the users • Reduce the probability of generating disinformation and misinformation.
3. App that rewards its users for flagging issues	This app rewards its users through monetary gains or other gifts to promote the flaggin of issues in chatGPT. This enhances the probability that issues get flagged by its users rather than be ignored and go unnoticed	• Prevents misinformation from reaching the users entirely.
4. Data Sanitiser	A service website that accepts text data as an input and outputs sanitised data (i.e. irrelevant, inaccurate or false information removed).	• Adding an additional layer of filtering to ChatGPT's training dataset reduces the amount of misinformation it can provide.
5. Live monitoring of user prompts	Using third-party applications one could implement a live monitoring service that flags and intercepts user prompts that indicate misuse/bypass attempts of the system. Implement a timeout/ban system for frequent offenders (i.e. getting flagged too frequently)	• Regulate the misuse of ChatGPT much more strictly • Reduce the probability of generating problematic outputs (fabricated articles/news, malware, etc.)
6. Implement Paywalls	Erect a paywall before providing responses to illegal or unethical user queries. Users are still able to obtain a censored version for the free version.	• Limits the accessibility of using ChatGPT for disinformation
7. Fake News Detector	Training is an AI model that can detect false information on social media based on user posts and comments.	• Reduces the spread of misinformation or use of disinformation on the internet. • Counters the use of ChatGPT for disinformation purposes.
8. Restrict User Input	Train an AI model that can detect keywords that have a high chance of outputting problematic answers.	• Can preemptively avoid situations where disinformation is generated

	If this happens ChatGPT will simply not answer the question	by ChatGPT

2.1.1.6 Deliberate

After holding a majority vote, we decided to use the idea of setting up a Forum Website.We then deliberated our proposed solution based on eight points of view, which are

Human, societal and environmental well-being: The website aims to let developers quickly get hold of information on vulnerabilities present in ChatGPT and fix them as soon as possible. This benefits human society as a whole since it reduces the chances of ChatGPT being exploited for disinformation.

Human-centered values (human rights, human intellect and dignity, diversity, and the autonomy of individuals): The website ensures that we are able to take the steps to abide by human-centered values. It will serve as a platform that lets users address nuanced issues within our society that ChatGPT was unable to identify.

Fairness: The website offers all users the same right to report questions whether you are a ChatGPT plus user or not. Users' posts are prioritised based on the time of post, and not based on their subscription to ChatGPT.

Privacy protection and security: This website allows users to create posts in public or private mode. Users who don't want to reveal sensitive information can opt for creating a private post instead. Forum posts will go through some basic form of sanitisation to prevent cross-site scripting attacks.

Reliability and safety: This website adopts multi-party discussion on user-reported issues, which effectively decentralizes OpenAI's monopoly over user reports, where they can no longer selectively answer user reports for their own benefit.

Transparency and explainability: Developers have a channel to share information about updates such as security patches, bug fixes etc. Which is originally hard to navigate to.

Contestability: Quoted from the textbook, "when an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or outcomes of the AI system." The website allows users to raise possible misinformation encounters when using ChatGPT and challenge the response given using the website,

Accountability: Developers have a specified tag with their accounts for other users to identify their field of work and their responsibilities in creating ChatGPT.

2.1.1.7 Produce solution

With all the information from our research and discussions, we have sufficient knowledge about the topic at hand, which are the problem's influence factors, ACM Code violations, and potential solutions. Using ethical theories as a guideline, we formulate viable ideas that may resolve the problem, or reduce the impact of the problem onto its various stakeholders.

The result of the deliberation phase further boosts our confidence that our proposed solution is in line with ethical principles and IT professionalism. We are able to deliver the proposal of the solution as written in section 3 of this report.

2.2 Techniques/Methods for Teamwork

2.2.1 Techniques

These are the techniques that we used throughout the assignment to conduct research, think of ideas, or resolve conflicts within the team.

Majority Voting : All voting sessions are conducted using majority voting where the majority voted item will be the final result. This is to facilitate selection processes as it is a straightforward way to arrive at a decision without extensive debate or discussion.

Teaming : Splitting up into pairs to conduct research and present findings during meetings. We think that by working in smaller groups, research efficiency will be greatly improved because the pair will have a more focused point of research compared to researching a more general topic as a big team.

Brainwriting : Group members independently write down ideas either synchronously or asynchronously. Those ideas are then gathered, shared, and discussed in a group. This is used to identify potential factors and ACM code violations without the issue of group members influencing one another.

Brainstorming : We as a group generate ideas freely and openly. These sessions do not involve the evaluation of ideas, we kept a clear head of the goal, which is to generate as many ideas as possible. We think that this is the most effective method for ideation since the deliberation phase is not included in this process.

Workable compromise : In the case where arguments arise, we sought for a compromise between the two differing points of view. This is mostly applied when group members have differing schedules due to different units' assignments, where we have to decide on a deadline for the action items that is fair for everyone.

Set common goals, expectations and schedule : We had a meeting to agree upon the common goals and expectations for this assignment, as well as set the schedule for when to conduct research and hold meetings. This is to ensure that every member of the team is always on the same page, and that we can deliver on time during each meeting.

2.2.2 Tools

These are the tools that we used mainly for effective communication and collaboration when carrying out tasks of the assignment such as writing the report and holding meetings.

Google Docs : A tool which users can create and edit documents online while collaborating in real-time with others. We used this to write the report and share meeting minutes with one another..

Zoom : A video conferencing platform that allows people to meet virtually from anywhere in the world. It also has useful tools such as polling which we used for voting. We think that online meetings have higher efficiency than physical meetings, along with lower risk of exposure to the current ongoing COVID-19 pandemic.

WhatsApp : A popular instant messaging application that allows users to send text messages, voice messages, make voice and video calls, share images, videos, and documents with other WhatsApp users in real-time. We used this app as a main platform for communication, e.g. scheduling meetings, sending notifications, sharing documents and sometimes reminding people to not miss meetings.

2.2.3 Tuckman's Stages of Team Development

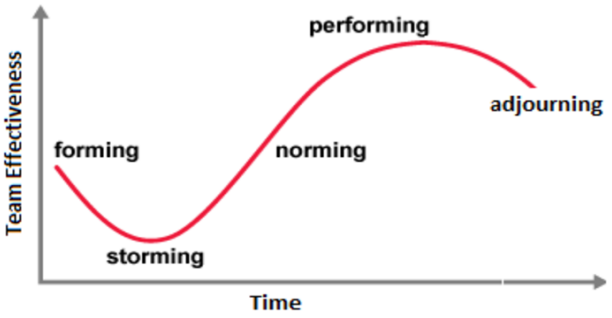

Figure: Stages of Tuckman's Team Development Model W.R.T. Time and Team Effectiveness (Lumen Candela, n.d.)

Tuckman's Team Development Model helps us understand the stages of development a team goes through at work, there are a total of five stages, namely Forming, Storming, Norming, Performing and Adjourning. However, the last step, which is Adjourning, is excluded since our team composition remains the same for the next assignment.

Forming is the stage when the team is first formed. Our team members already had several interactions in the previous weeks' tutorials, therefore we already knew about each other. This allowed us to dive right into the discussion of laying down ground rules, forming the team contract, and delegating roles. The effectiveness of the team is quite satisfactory as everyone has similar understandings on the dos and don'ts when it comes to working as a team.

Storming is the period marked by conflict and competition as individual personalities emerge. (Lumen Candela, n.d.) However, our team quickly transitioned through this phase because our members always state a workable compromise every time they raise an objection. The suggestion can be readily accepted by both parties and our meetings can continue without much disruption.

Norming is the phase where conflict is resolved and some form of unity is established in the team. Team performance increases during this stage where group members were less focused on interpersonal differences, and started cooperating to work towards the team goal.

Performing is the stage where all team members come to the same consensus where they fully respect each other, and team performance is at peak. All members of the team accepted the team leader’s task delegation without complaints, completed their action items responsibly, and always delivered great results during each meeting.

3. Proposed solution

In the ever-evolving landscape of artificial intelligence and natural language processing, ChatGPT has emerged as a remarkable tool that can engage in dynamic and informative conversations with users. As ChatGPT continues to advance, the collaboration between developers and users becomes increasingly crucial in refining its capabilities and addressing potential issues. To foster this symbiotic relationship, a novel approach has been introduced – a dedicated website that enables users to directly interact with ChatGPT developers, empowering them to uncover, report, and address any problems they encounter.

Key features:

1. Users are able to create public or private threads (useful when flagging exploits)

Users are given the flexibility of creating either public or private threads within the interactive website, allowing for diverse and personalized interactions. Whether seeking assistance, sharing experiences, or discussing specific issues, users can opt for public threads to engage in open conversations with the broader user community, fostering knowledge exchange and collaborative problem-solving. Conversely, private threads provide a secure and confidential space for users to communicate directly with ChatGPT developers, enabling them to address sensitive concerns or provide detailed feedback privately.

The public and private thread feature is designed to promote responsible and secure sharing of information on the feedback website. By allowing users to create private threads for discussing sensitive information, this can help[ensure that potential exploits or vulnerabilities are reported and addressed in a responsible and secure manner.

2. Users are given a credibility system, where users can rate the relativity of a post, the helpfulness of a comment, and the trustworthiness of another user.

To add on to the user threads, the website will also facilitate a reliable and effective user feedback system. By providing the option to rate the relativity of a post and the helpfulness of comments, users can contribute to filtering and highlighting valuable content, ensuring that the most pertinent information receives prominence. Additionally, the ability to rate the trustworthiness of other users fosters a sense of accountability and transparency within the community, enabling users to identify credible sources and engage in more meaningful discussions. The credibility system is designed to promote responsible and trustworthy behavior on the feedback website, and to help users make informed decisions about the content they consume and the users they interact with. By providing users with a way to evaluate the relativity, helpfulness, and trustworthiness of posts, comments, and the other users, the system can help create a more transparent and trustworthy community of users.

3. Developers have a dedicated page for patch note releases

By allowing users to engage in conversations with developers, it not only facilitates issue identification but also fosters a sense of community and collaboration. This proactive involvement of users ultimately leads to the improvement of ChatGPT's performance,

usability, and overall user experience. The platform also allows OpenAI and their developers to be more transparent with their user base, regarding issues addressed by their users, bugs and fixes, patch notes, etc. Transparency is a key factor that helps build trust between its users, which is a major foundation of effective collaboration and teamwork. Additionally, mutual trust can be beneficial in promoting active participation in their forums.

4. Developers are given a tag to indicate their responsibilities, as well as credit their contribution in developing ChatGPT.

These tags will help users to identify who to reach out to when addressing issues on the site. This will ensure that the problems raised by users will be handled appropriately and efficiently.

5. A search bar for users to effectively search for topics.

Users are able to access the search bar from the main page of the website and can be used to quickly and easily find relevant threads, posts, and comments related to specific topics.

The search bar is designed to promote efficient and effective problem reporting on the feedback website. By allowing users to quickly find relevant information related to their problem, this can help users address issues more quickly and effectively. Additionally, this feature can help reduce duplicate problem reports by directing users to existing threads or posts related to their issues

6. A filter system with tags and categories to filter for appropriate forums to respond to.

The system allows the users to filter for threads via specific words (a.k.a. tags) or via categories. This is to facilitate user navigation around the platform and effectively find posts of interest.

7. Recommendation system

The website is able to provide recommendations to the user for posts relevant to what they searched for. The system is also able to recommend currently popular posts to anonymous users. The recommendation system displays a list of the most popular or trending topics on the website, as determined by the number of posts, comments, or ratings related to each topic. Users can click on a topic to view related threads, posts, and comments, allowing them to quickly find information related to the topic.

8. Moderation tools and Moderators

Moderation tools are essential in any forum website, enabling administrators to manage and moderate threads, delete inappropriate content, ban users and move posts to appropriate categories.

9. FAQ section

The Frequently Asked Questions section provides users with a list of commonly asked questions and answers. Users can browse the questions and answers to find information related to their issue, or they can search the section using the search bar feature to find specific topics. Users can quickly and easily find relevant information related to their issue or question through the FAQ section.

4. Conclusion

To conclude, the ethical issue pertaining to Generative AI, specifically ChatGPT, is its ability to generate misinformation and disinformation. Factors such as no self-validation algorithms within ChatGPT, the usage of large datasets that are prone to inaccuracies and biases, and the susceptibility to policy-bypassing prompt injection are believed to cause such an issue. The problem is narrowed in relation to the ACM Code of Ethics, we identified the violation of three main codes which are Code 1.2 Avoid Harm, Code 1.4 Be fair and take action not to discriminate, and Code 2.5 Give comprehensive and thorough evaluations of computer systems and their impacts. We also had a discussion on the ethical theories and laid down general principles that serve as a guideline for ideating possible solutions. We selected the idea of implementing a User Forum in the form of a website as our proposed solution.

The User Forum website has several notable features, such as the ability to create public and private threads, a user credibility system, a dedicated page for patch note releases, developer tags, a search bar, a filter system using keyword tags and categories, a recommendation system, moderation tools, and a FAQ section which provides users a list of commonly asked questions and solutions. The private thread feature protects user privacy such that they will not unintentionally publicize sensitive information. Moderation tools and moderators are to enforce a basic level of security on the site to prevent inappropriate posts and misuse of the forum. The user credibility system motivates users to actively engage in the forum and provide helpful comments or create impactful discussion topics regarding ChatGPT thus ensuring the site will have a constant positive significant impact on resolving ChatGPT related issues. Having a feature that allows the developers to communicate their decision-making and fixes helps in increasing transparency, helping to build mutual trust between OpenAI and the user base as a whole, which can be beneficial for future developments. The search feature, recommendation system, and FAQ section provides a user-friendly interface that allows users to efficiently navigate through the website.

Overall, the forum website can be an invaluable tool for improving the performance of ChatGPT and promoting responsible and trustworthy behavior among users. It can also tackle the ethical issue of misinformation and disinformation with the collective efforts of its users and developers. With careful monitoring and maintenance, developers can ensure that it remains a valuable resource for users and a key component of the ChatGPT ecosystem.

References

Dehouche N. (2021) “Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3)”

Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3) (intres.com)

Association for Computing Machinery, 2018 ACM Code of Ethics and Professional Conduct

Layout 1 (acm.org)

Lumen Candela (n.d.) “The Five Stages of Team Development

The Five Stages of Team Development | Principles of Management (lumenlearning.com)

Nailah. (2022). Moodle Monash. FIT1055 Textbook

c45d0ebf4f1411d12390ae8c670f830e1265f93e (d3cgwrxphz0fqu.cloudfront.net)