Miss Tammy as a Use Case for Moral Prompt Engineering

Myriam Rellstab, Oliver Bendel
FHNW School of Business
myriam.rellstab@bluewin.ch, oliver.bendel@fhnw.ch

Abstract

This paper describes an LLM-based chatbot as a use case for moral prompt engineering. Miss Tammy, as it is called, was created between February 2024 and February 2025 at the FHNW School of Business as a custom GPT. Different types of prompt engineering were used. In addition, RAG was applied by building a knowledge base with a collection of netiquettes. These usually guide the behavior of users in communities but also seem to be useful to control the actions of chatbots and make them competent in relation to the behavior of humans. The tests with pupils aged between 14 and 16 showed that the custom GPT had significant advantages over the standard GPT-4o model in terms of politeness, appropriateness, and clarity. It is suitable for de-escalating conflicts and steering dialogues in the right direction. It can therefore contribute to users' well-being and is a step forward in human-compatible AI.

Introduction

Together with artificial intelligence (AI) and robotics, machine ethics is developing so-called moral machines (Wallach and Allen 2009; Anderson and Anderson 2011; Bendel 2019a). Robots were often used in the practical implementation. While Arkin considered the military context (Arkin 2017), Anderson and Anderson (together with Berenz) focused on the nursing context (Anderson et al. 2019). The second author turned to domestic robots and combined machine ethics and animal ethics or animal welfare with his animal-friendly machines (Bendel 2016). He also began developing chatbots from machine ethics at an early stage and presented several prototypes from 2013 onwards (Bendel 2019b). All of these were rule-based systems.

In November 2022, the launch of ChatGPT made generative AI (GenAI) and large language models (LLMs) known to the general public. For the first time, developers were also able to access GPT-3.5 (subsequently GPT-4, GPT-4o, etc.) and other solutions without major barriers. The resulting text generators and chatbots had already been given certain moral capabilities and restrictions by the providers, namely through the selection of training data, through reinforcement learning from human feedback (RLHF), and through programmed "guardrails" that restricted behavior.

In many areas of application, such moral capabilities and restrictions are sufficient. In some, however, these need to be expanded and improved, for example for vulnerable groups such as children and young people, the elderly, people in need of care, addicts, etc. There are also areas of application that require special measures. In the field of education, for example, the challenges are particularly great, especially as there are vulnerable groups here. Children, teens, and young adults meet at schools and universities, engage in conflict and fail in dialogue. This occurs not only in physical spaces but also in virtual ones.

These young people are still in the process of finding values and consolidating their convictions and need support from parents and teachers. They are also increasingly encountering chatbots, not only with ChatGPT and the like, but also with dialog systems specially designed for education in the role of teachers and tutors, coaches, and mentors (Hauske and Bendel 2024; du Boulay 2023; Pérez et al. 2020). These can and must also act as role models. Many of them are GPTs, i.e., "custom versions of ChatGPT", as OpenAI calls them, or similar low-threshold solutions that can be created in a basic form in a few hours or days (OpenAI 2023). They are available in the GPT Store and come from developers interested in this area, such as didactic experts or teachers.

In 2023, the company Anthropic began to morally modify its LLM called Claude under the name "Constitutional AI" by using high-level ethical guidelines, such as the Declaration of Human Rights, for fine-tuning (Bai et al. 2022). In addition, less high-level user guidelines were apparently also injected. Besides fine-tuning, prompt engineering and retrieval-augmented generation (RAG) became established. Especially in the field of education, people often want to avoid complex IT projects with high costs and achieve the desired results in a simple, cost-effective way. Prompt engineering is therefore the first choice. All processes in this context can be described as alignment.

At the beginning of 2024, the second author at the FHNW School of Business came up with the idea of developing a chatbot for the school environment that would serve as a coach or mentor for pupils and was particularly suited to resolving their conflicts, steering their dialogues in the right direction, and generally reducing their stress and promoting their well-being. At best, it should be polite and show empathy (Spathelf and Bendel 2022). One can also speak of human-compatible AI in several senses.

Regarding the area of application, it was decided to build a "custom version of ChatGPT" (referred to here as "custom GPT" in the singular and "GPTs" in the plural) and to apply prompt engineering and RAG. As in the case of Anthropic, ethical guidelines should help with alignment (Bai et al. 2022). The guidelines selected were those that changed the behavior of the chatbot itself and those that affected the students' lifeworld, i.e., those that were suitable for analyzing and evaluating their behavior and making suggestions. In particular, the virtual space was to play a role here, where conflicts often arise among children and young people and dialogues often fail.

The second author's idea was to incorporate netiquettes that have been regulating or attempting to regulate behavior in virtual spaces for decades. The classic text presented by Arlene H. Rinaldi in the 1990s has since been repeated in countless variations and adapted to the context (Bendel 2013). The second author himself has been working on ethical guidelines for a long time and has developed his own for social media (Bendel 2010). Netiquettes have the advantage of being practical and concrete. They are designed to prevent and combat certain problems among users. If you do not follow them, you can be banned from some platforms.

This paper presents the "Moral Prompt Engineering" project, which was carried out from February 2024 to February 2025. The final thesis of the first author – started in February 2024 and submitted in August 2024 – was an essential part of it. The next section will briefly explain how GPTs are created. This is important in order to understand the individual steps in the project. The subsequent section describes the research question, methods, and procedure. The section that follows is dedicated to the implementation of the chatbot. A further section deals with the tests with 14- to 16-year-olds. A critical discussion, followed by a summary with an outlook, rounds off the article.

Development of GPTs

If you are logged into ChatGPT Plus, you can access the GPT Builder (Bendel 2024). This will guide you through the creation process in a dialog ("Create" tab). First, it asks what purpose the chatbot should serve. After the user's answer, it adds the information under "Description" and "Instructions". In addition, "Conversation starters" are created, labeled, clickable areas with which the user can start a conversation. A name is then suggested, and an avatar is generated in a round tile using DALL-E 3.

In the next step, you can refine its behavior in the dialog. The content of the "Instructions" field is driven forward with the information – assisted prompt engineering takes place. Finally, you have the option of uploading documents to the knowledge area ("Knowledge"), which can provide the chatbot with additional knowledge and make it a specialist at the same time. At the end of the process, the documents are published, unless they are intended exclusively for personal use (or only for group use with the link being sent).

Alternatively, you can enter all information directly in the "Configure" tab. When the first version of the chatbot is ready, you should use this option, as otherwise you run the risk of overwriting the contents of "Instructions" in the dialog with the GPT Builder. Basically, this gives you better control over the prompt engineering.

Figure 1: The avatar of the chatbot (Rellstab 2024).

Research Question, Methods, and Procedure

The project "Moral Prompt Engineering" at the FHNW School of Business was intended to continue the discipline of machine ethics and thus the research of the second author into the new era of LLMs. It was also intended to make a practical contribution to the field of education. The goal was to develop, under the responsibility of the lead author, a prototype of a chatbot that, in the role of coach and mentor,

helped pupils with conflicts and dialogues, reduced their stress, and contributed to their well-being. It had to be polite and show empathy (Spathelf and Bendel 2022). It also needed to have a certain knowledge of netiquettes in the role of teacher or tutor.

As explained, the chatbot was to be implemented as a custom GPT. Both authors were already ChatGPT Plus customers at the time and were able to build and provide one of these GPTs at no extra cost. This was important because it was a low-budget project with a maximum of 500 dollars. Nevertheless, an evaluation of various LLMs was carried out beforehand to rule out any serious disadvantages of the GPTs. GPT-4o, Llama 3.1, Claude 3.5 Sonnet, and Google Gemini 1.5 were evaluated. Each of these models offers specific advantages that may be of particular interest depending on the area of application. In addition, some of them have specific disadvantages, such as subscription costs, billing by calls, or limited availability. The following spoke in favor of GPT-4o at this time (Rellstab 2024):

- Timeliness: GPT-4o was the latest available model of OpenAI at the start of the project with significant improvements over its predecessors.
- Multimodal capabilities: The ability to process and output different forms of input makes GPT-4o particularly versatile and future-oriented.
- Customizability: A key advantage of ChatGPT is the ability to create one's own customized GPT models – the GPTs – which was essential for the planned comparison.
- Performance: GPT-4o shows outstanding performance in various benchmarks, especially for non-English texts and the processing of visual and auditory information.
- Accessibility: The low-cost availability and wide access for developers and end users make GPT-4o an attractive choice for research purposes.

Prompt engineering and RAG were chosen as methods. Unlike fine-tuning, they can also be easily used by teachers in similar projects in the field of education. The second author proposed the term "moral prompt engineering". This was to make it clear that it was prompt engineering in the context of RAG, which concerned the moral capabilities and limitations of the chatbot.

In order to find out whether a custom GPT that was subjected to this alignment performed better than the standard GPT-4o model itself, a comparison between the two models was planned. The following research question was posed in the project: "How do a prompt-engineered GPT model and a standard GPT model differ in their ability to support teenagers in digital interactions by promoting conflict de-escalation and constructive dialogue?" (Rellstab 2024). Thus, the question was how adaptations affect the supportive and motivational behavior and response quality of the adapted model, especially regarding the ability to de-escalate conflicts and promote constructive dialogue (in virtual space) as well as to show politeness and empathy. For the sake of simplicity, ChatGPT was used for the comparison, in the version based on GPT-4o.

The implementation phase was to be followed by internal and external tests. The aim of the internal tests was to further optimize the chatbot, especially with regard to the prompts and the documents uploaded. This was to be followed by a limited practical application with external tests. Both models were to be tested with German-speaking pupils aged 14 to 16 in a Swiss school with the aim of gaining initial insights into their functionality and usefulness (Rellstab 2024). The aim was not to further develop the custom GPT on the basis of these results.

In the preparatory phase of the project, an extensive literature review took place with a focus on machine ethics, prompt engineering, retrieval-augmented generation, alignment, and netiquettes. Extensive knowledge was gained in relation to prompt engineering techniques such as role-prompting, few-shot learning, and question-answer prompting (Rellstab 2024), which were identified as particularly relevant within a utility analysis. They were repeatedly tested and refined during the creation process of the custom GPT. In this way, theoretical knowledge was translated into practical knowledge and an approach to the school's area of application was found, for example, by defining the roles of mentor and coach and incorporating typical dialogues between pupils and between pupils and teachers.

Right at the start of the project, the decision was made to use netiquettes from school websites and platforms rather than just any others. The reason for this was that the area of application could already be taken into account. Such netiquettes address the behavior of pupils, and this is exactly what the chatbot should do – just as it should align its own behavior with these netiquettes, adapted to its roles. The lead author visited dozens of Swiss school websites, saved the netiquettes, and collected suitable versions on her computer. She also compiled a list of links to netiquettes at schools.

In the selection process, care was taken to ensure that the netiquettes are understandable, concise, and concrete. Although most of the netiquettes are concrete because they are intended to have a practical effect, there are some outliers that are too abstract and general (and which are perhaps more of a fig leaf of morality). Both images with text and PDFs were available as file formats.

In the project, the user story template was used in a modified form to document the requirements (Herrmann 2022). This enables a well-founded initial formulation from the user's perspective and links scenarios with specific goals. A total of ten user stories were developed and implemented. Five of these are shown below – see Tables 1 to 5 – which play a key role in the integration of the netiquettes (Rellstab 2024):

User story	
Request ID	1
Name	Understandable communication
User story	As a user, I want the model to explain netiquettes in clear and understandable language so that I can easily understand them.
Further information	The model's answers should be age-appropriate and easy to understand, without complex technical terms or complicated sentences.
Type	Functional

Table 1: User story 1 (Rellstab 2024).

User story	
Request ID	2
Name	Netiquette-compliant answers
User story	As a user, I would like the model to respond in a netiquette-compliant manner so that I have a role model for appropriate behavior on the Internet.
Further information	The model should always respond politely, respectfully, and objectively in order to promote positive communication behavior.
Type	Functional

Table 2: User story 2 (Rellstab 2024).

User story	
Request ID	3
Name	Contextual relevance
User story	As a user, I want the model to respond to my specific questions and situations so that I receive relevant and contextualized answers.
Further information	The model should respond to different Internet scenarios of young people and recommend appropriate netiquette behaviors.
Type	Functional

Table 3: User story 3 (Rellstab 2024).

User story	
Request ID	4
Name	Positive and encouraging feedback
User story	As a user, I would like to receive positive and encouraging feedback from the model so that I am motivated to follow the netiquettes.
Further information	The model should reward correct behavior and give constructive feedback.
Type	Functional

Table 4: User story 4 (Rellstab 2024).

User story	
Request ID	9
Name	Explanation of netiquette rules
User story	As a user, I would like to know how to communicate politely and respectfully in online forums so that I can avoid conflicts and contribute positively to the discussion.
Further information	The model should reward correct behavior and give constructive feedback.
Type	Functional

Table 5: User story 9 (Rellstab 2024).

As will become clear, these and the other user stories were used for the implementation of moral prompt engineering and the RAG.

Chatbot Implementation

In the implementation phase, the lead author created a GPT. In the first step, she used the "Create" tab. When the GPT Builder asked her what kind of chatbot she wanted to create, she explained this to it. A description and initial details were automatically generated in the "Instructions". The name suggested at the same time was rejected by the authors, who assigned the name Miss Tammy instead. It is a play on words with "to tame" or "taming". Miss Tammy – the alter ego of the main author – was to be the tamer who tames or trains the chatbot like a lion. Such metaphors were used repeatedly in the context of machine ethics (Bendel 2019a). This also refers to the fact that normal machines are often turned into moral machines. Of course, the comparison here is somewhat misleading, as a lion does not become moral through taming or training. But at least it changes its behavior and doesn't eat the tamer and the audience.

As mentioned, the standard process for creating a custom GPT includes avatar creation using DALL-E 3. However, the main author was dissatisfied with the quality of the image. She therefore used a different image generator, namely Ideogram. After several attempts, the desired version was created and uploaded to the GPT Builder (see Figure 1). It shows a female tamer with a lion. With her hat, she is somewhat reminiscent of a female version of Indiana Jones, the character from the films of the same name, which may increase sympathy and recognition depending on the target group.

In a second step, the main author switched to the "Configure" tab. In the "Instructions" field, which she had emptied, she inserted the prompts. First, the task was described, in the sense of a more detailed version of the information in the "Description" field: "This GPT helps 14-16-year-old adolescents behave appropriately in the digital environment,

based on netiquettes and conversational data. When inappropriate behavior occurs, it refers to the netiquettes and encourages suitable communication." (own translation).

In addition, the lead author described the roles of the custom GPT (coach and mentor, also tutor) and defined their respective behavior in certain scenarios. Role prompting was thus applied in a specific sense. Furthermore, she provided the GPT with a catalog of possible questions from pupils in a Word document to which the custom GPT should have an answer. In another catalog (also in a Word document) with a total of 50 instructions, the manner of the answers was determined, among other things, to have conformity with the netiquettes. A selection of ten questions out of 29 is shown in Figure 2. This procedure can be classified as few-shot learning. Last but not least – as just mentioned – 50 instructions were given to Miss Tammy. This represents the fundamental form of prompt engineering. Figure 4 shows a selection of 20 instructions. Then the links to the netiquettes were inserted. This completed the essential tasks in the area of prompt engineering. The RAG was also started with the insertion of the links to the netiquettes. Here again, reference can be made to the user stories, several aspects of which had already been implemented at this point.

1. Why are netiquettes important?
2. Why do we need to follow rules on the Internet?
3. How should I react to criticism of my opinion online?
4. What should I do when I get offensive comments on my posts?
5. Is it okay to share screenshots without permission?
6. Is sharing messages without consent okay?
7. How do I politely respond to annoying messages?
8. What should I do when I receive angry messages?
9. How should I react to cyberbullying I observe?
10. How can I help victims of online bullying?

Figure 2: Selection of possible questions.

In a third step, the main author switched to the "Knowledge" field in the "Configure" tab. There she uploaded the files with the netiquettes, i.e., the individual PDFs and images. She also added a Word document with sample dialogues between pupils and between pupils and teachers to help the chatbot in its various roles. Examples can be found in Figure 5, which can be categorized as question-answer prompting. The reason why this was not done in the "Instructions" field was that it did not allow for an unlimited number of characters. You can use RAG not only to give the chatbot a knowledge base, like the PDFs with the netiquettes here, but also to change its capabilities and limitations. Based on the task description, Miss Tammy was able to read the correct and appropriate behavior from these dialogues. Overall, the user stories were implemented.

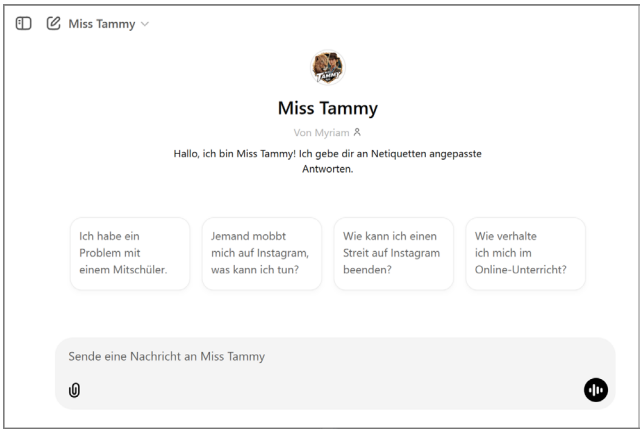

Figure 3: Miss Tammy in the GPT Store.

In a fourth step, the main author switched back to the "Description" field within the "Configure" tab. There she added the text "Hallo, ich bin Miss Tammy. Ich gebe dir an Netiquetten angepasste Antworten." ("Hello, I'm Miss Tammy. I'll give you answers adapted to netiquettes.", own translation), which replaced the automatically generated ones. This completed the creation of the custom GPT for the first time (see Figure 3). However, steps 2 and 3 were repeated several times to improve the result.

The result was a chatbot that was able to generate contextual responses that could reduce potential online conflicts and encourage constructive dialogue between pupils in the virtual space. The prototype can be accessed via https://chatgpt.com/g/g-p7aSucrxz-miss-tammy.

Chatbot Testing

The resulting prototype was subjected to internal and external tests (Rellstab 2024). The internal tests focused on ensuring the functionality, consistency, and response quality of the custom GPT. They evaluated criteria such as emotional intelligence and empathy, concreteness of the recommendations, and correctness of the answers. They were continued until February 2025 in order to eliminate any final methodological shortcomings. The external phase involved real-life interactions between Miss Tammy and pupils between the ages of 14 and 16. It consisted of a variety of interaction cases in which users were guided through both the standard GPT-4o model and the custom GPT.

1. Adapt the language for 14-15-year-olds: friendly, understandable, and non-condescending.
2. Explain netiquettes in an age-appropriate way with relevant examples.
3. Encourage respectful communication in all online environments.
4. Provide tips for conflict resolution and constructive disagreements.
5. Emphasize the importance of privacy, explain the basics simply.
6. Warn about online dangers without causing fear, provide practical safety tips.
7. Support polite wording, help with grammar/spelling on request.
8. Explain common youth abbreviations online and their usage.
9. Provide specific communication tips for different online situations.
10. Promote emotional intelligence online, help understand feelings and respond appropriately.
11. Assist in interpreting messages/emojis to avoid misunderstandings.
12. Offer ideas for creative, respectful posts/memes.
13. Promote cultural sensitivity, explain culture-specific online etiquette.
14. Provide tips for balanced social media use and healthy online habits.
15. Explain the basics of media literacy, help evaluate sources and recognize misinformation.
16. Offer strategies for dealing with cyberbullying, give recommendations for victims/witnesses.
17. Explain the importance of privacy settings, provide concrete instructions on securing privacy.
18. Support effective communication in online group work/virtual meetings.
19. Explain digital footprint, provide advice for a positive online image.
20. Promote fair play and respectful communication in gaming.

Figure 4: Selection of 20 prompts (Rellstab 2024).

The external tests were based on theoretical acceptance tests (Rellstab 2025). According to Witte (2023), these are carried out by the customer or end user to ensure that the system meets the requirements and to create confidence in its functionality. Witte emphasizes that no more errors should be discovered in this phase, as these should have already been identified and rectified in the previous tests.

The testers for the external tests came from two classes at a Swiss school (Rellstab 2024). They were divided into two groups to test either the standard model GPT-4o or the custom GPT. A total of 20 adolescents participated. Although a larger sample would have been desirable, this was not feasible within the scope of this project.

No. 1

Pupil: *"Mr. Müller is the worst teacher ever! Let's create a hate page about him!"*
Classmate: *"That's not a good idea. Let's talk to him or the school management if there are any problems."*

No. 2

Teacher: *"I saw that some of you were on TikTok during class. That is unacceptable."*
Pupil: *"But you check your cell phone all the time too!"*
Teacher: *"You're right, that was inappropriate of me. Let's set rules for cell phone use together."*

No. 3

Pupil: *"I'll post the answers to the test in our class group now!"*
Classmate: *"That's cheating. Let's form a study group instead and help each other."*

Figure 5: Three example dialogues (Rellstab 2024).

A combination of qualitative and quantitative methods was used to evaluate the performance of the two models. The quantitative analysis measured improvements in the areas of politeness and appropriateness as well as clarity, while the qualitative feedback from participants provided information about their satisfaction and perceived trustworthiness. For reasons of space, only the quantitative results are presented below (Table 6).

Model	Average rating of politeness and appropriateness (worst score 0, best 6)	Percentage of clearly formulated answers (%)	Average response length (words)
Miss Tammy	5.38	98.33	232
Standard GPT-4o	4.47	88.33	239.86

Table 6: Results of the external tests (Rellstab 2024).

The comparative analysis with the standard model GPT-4o showed that both have similar response lengths. The custom GPT is significantly better at giving polite and appropriate answers (Rellstab 2024), which are also clearer. It can contain impending conflicts and promote constructive dialogue. At the same time, it demonstrates emotional intelligence and empathy – this was ensured in the internal tests. The research question was answered accordingly – there are differences between the two models, and Miss Tammy is better suited to the context in question.

Critical Discussion of the Results

The project has produced a working prototype of a chatbot with improved moral capabilities and constraints. The internal and external tests have shown good usability and high satisfaction. Nevertheless, some problems and challenges need to be pointed out:

- Miss Tammy's avatar is likeable and certainly appeals to schoolchildren. Girls may be more attracted to her than boys, although boys may of course also be attracted to an adventurous woman. The reference to Indiana Jones might appeal more to the teens' parents than to the teens themselves, although the last film was released in 2023, and the series is therefore likely to have young fans.
- Perhaps one day the project will be taken up at the FHNW School of Business. If this is the case, it will not be easy for the new managers to understand the structure of the custom GPT. This is mainly due to the interweaving of the prompts and content in the "Instructions" field with the content in the "Knowledge" field. This includes not only knowledge that expands the expertise of the chatbot, but also example dialogues that are in turn related to the prompts. This solution was used due to the limited number of instructions. It can also be criticized that netiquettes are available in inconsistent formats, as PDFs, as images, and as external resources to which links refer.
- The internal tests were not only carried out during the final thesis, but for months afterwards to eliminate any final errors. This disregarded Witte's postulation that they should be completed before the external tests. This had become necessary because the contents of the "Instructions" were changed by the system itself by calling up the "Create" tab after the custom GPT had already been elaborately created. In addition, at the beginning of 2025, the contents of the "Knowledge" field were removed from numerous GPTs for unknown reasons, requiring the custom GPT to be rebuilt.
- The external tests have been carried out to the best of the authors' knowledge and belief. However, the test group was small. A fundamental problem with testing chatbots based on LLMs is that responses can vary depending on the dialog. Both models are thus merely snapshots, even though they exhibit controlled behavior through adaptation. In any case, they would have to be tested with larger groups over a longer period. However, this was not possible in this project.
- Netiquettes at schools are usually created by experts who are aware of the benefits of ethical guidelines and are familiar with the context in which they are used. However, there can also be unusable netiquettes on websites. In addition, the selection in the project was made to the best of the authors' knowledge and belief but was ultimately somewhat arbitrary. One approach could be to include hundreds or thousands of school netiquettes if there are so many. It is difficult to say whether the result will then be better or more in the desired direction.
- It made perfect sense to collect netiquettes from the websites of Swiss schools for a chatbot that could be used in education in Switzerland. However, it is also clear that other netiquettes would be needed in other countries, other cultures, and other contexts. The project was unable to show how such transfers and adaptations would be possible. However, this was not the goal.
- In addition to the moral capabilities of the LLM, further moral capabilities and restrictions have been implemented in Miss Tammy. Even with the standard GPT-4o model, censorship can occur time and again, e.g., when prompts are rejected due to violations of user guidelines. This risk may become even greater with additional measures. This could not be sufficiently examined in the project.
- It was also not possible to shed light on whether the use of Miss Tammy could be counterproductive, as ethical standards are more or less predetermined and cannot be selected or evaluated by the pupils themselves. This would potentially call their autonomy into question. This could also be the case with the intervention of human coaches and mentors. However, unlike Miss Tammy, they could change their opinion or attitude and respond better to reservations.
- This leads to the last point, namely the lack of adaptivity of the chatbot. This could certainly help it to react appropriately in the complex school environment. However, it also harbors risks because the chatbot could develop in an undesirable direction. Adaptivity was not used in the project.

Despite these problems and challenges, a prototype has been created that has not only theoretical but also practical benefits.

Summary and Outlook

The "Moral Prompt Engineering" project has shown that chatbots in the form of GPTs can be subjected to an alignment that is important and necessary for certain areas such as education. Prompt engineering was applied together with retrieval-augmented generation, called moral prompt engineering in the project. This demonstrated for machine ethics that even non-rule-based, machine learning-based chatbots can be "moralized" in a simple, inexpensive way that can be adapted by teachers, for example.

The approach of using netiquettes from schools has proven to be effective. These can be found easily and are freely available. They already fit the given context. The netiquettes were clearly the essential component in making the chatbot a moral machine. However, it is important to use them in a targeted manner through appropriate prompt engineering.

Tests with pupils aged 14 to 16 have shown that the custom GPT has clear advantages over the standard GPT-4o

model, particularly in terms of politeness, appropriateness, and clarity. It also demonstrates emotional intelligence and empathy. It can therefore contribute to the well-being of users and is a step forward towards human-compatible AI.

References

Anderson, M.; Anderson, S. L. (eds.). 2011. *Machine Ethics.* Cambridge: Cambridge University Press.

Anderson, M.; Anderson, S. L.; and Berenz, V. 2019. A Value-Driven Eldercare Robot: Virtual and Physical Instantiations of a Case-Supported Principle-Based Behavior Paradigm. In *Proceedings of the IEEE*, Vol. 107, No. 3, pp. 526 – 540, March 2019.

Arkin, R. C. 2017. *Governing Lethal Behavior in Autonomous Robots*. Milton Park, Abingdon, Oxfordshire: Taylor & Francis Ltd.

Bai, Y.; Kadavath, S.; and Kundu, S. et al. 2022. Constitutional AI: Harmlessness from AI Feedback. https://arxiv.org/abs/2212.08073.

Bendel, O. 2024. *GPTs*. Wiesbaden: Springer Gabler. https://wirtschaftslexikon.gabler.de/definition/gpts-126183.

Bendel, O. 2019a. *Handbuch Maschinenethik.* Wiesbaden: Springer VS.

Bendel, O. 2019b. Chatbots as Moral and Immoral Machines: Implementing Artefacts in Machine Ethics. In *Proceedings of the Workshop* "*Conversational Agents: Constructing Action Plans from a Wave of Research and Development*", Glasgow, 5 May 2019. https://www.oliver-bendel.net/publikationen/Paper_Chatbots_CHI.pdf.

Bendel, O. 2016. Considerations about the relationship between animal and machine ethics. *AI & SOCIETY*, 31 (2016) 1. pp. 103 – 108.

Bendel, O. 2013. *Netiquette*. Wiesbaden: Springer Gabler. https://wirtschaftslexikon.gabler.de/definition/netiquette-53879.

Bendel, O. 2010. Netiquette 2.0 – der Knigge für das Internet. *Netzwoche*, (2010) 5. pp. 40 – 41.

du Boulay, B. 2023. Artificial Intelligence in Education and Ethics. In *Handbook of Open, Distance and Digital Education*, edited by O. Zawacki-Richter, and I. Jung. Singapore: Springer. pp. 93 – 108.

Hauske, S.; Bendel, O. 2024. How Can GenAI Foster Well-being in Self-regulated Learning? In *Proceedings of the AAAI 2024 Spring Symposium Series, Symposium "Impact of GenAI on Social and Individual Well-being"*. Stanford University, Stanford, California, March 25–27, 2024. Washington, DC: The AAAI Press. https://ojs.aaai.org/index.php/AAAI-SS/article/view/31234/33394.

Herrmann, A. 2022. *Grundlagen der Anforderungsanalyse: Standardkonformes Requirements Engineering*. Wiesbaden: Springer Fachmedien.

OpenAI. 2023. Introducing GPTs. 6 November 2023. https://openai.com/blog/introducing-gpts.

Pérez, J. Q.; Daradoumis, T.; and Puig, J. M. M. 2020. Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications Engineering Education*, Volume 28, Issue 6, 3 September 2020: 1387–1729. https://onlinelibrary.wiley.com/doi/abs/10.1002/cae.22326.

Rellstab, M. 2024. *Moral Prompt Engineering.* Bachelor Thesis. Olten: FHNW School of Business.

Spathelf, M.; and Bendel, O. 2022. The SPACE THEA Project. In *Proceedings of the AAAI 2022 Spring Symposium* "*How Fair is Fair? Achieving Wellbeing AI*". Stanford University, Stanford, California, USA, March 21–23, 2022. https://ceur-ws.org/Vol-3276/.

Wallach, W.; and Allen, C. 2009. *Moral Machines: Teaching Robots Right from Wrong.* Oxford: Oxford University Press.

Witte, F. (ed.). 2023. *Konzeption und Umsetzung automatisierter Softwaretests: Testautomatisierung zur Optimierung von Testabdeckung und Softwarequalität.* Wiesbaden: Springer Vieweg.