Large-Scale Proliferation of Fake Persona Using Advanced Generative AI

Xiangyu Qi
Princeton University
xiangyuqi@princeton.edu

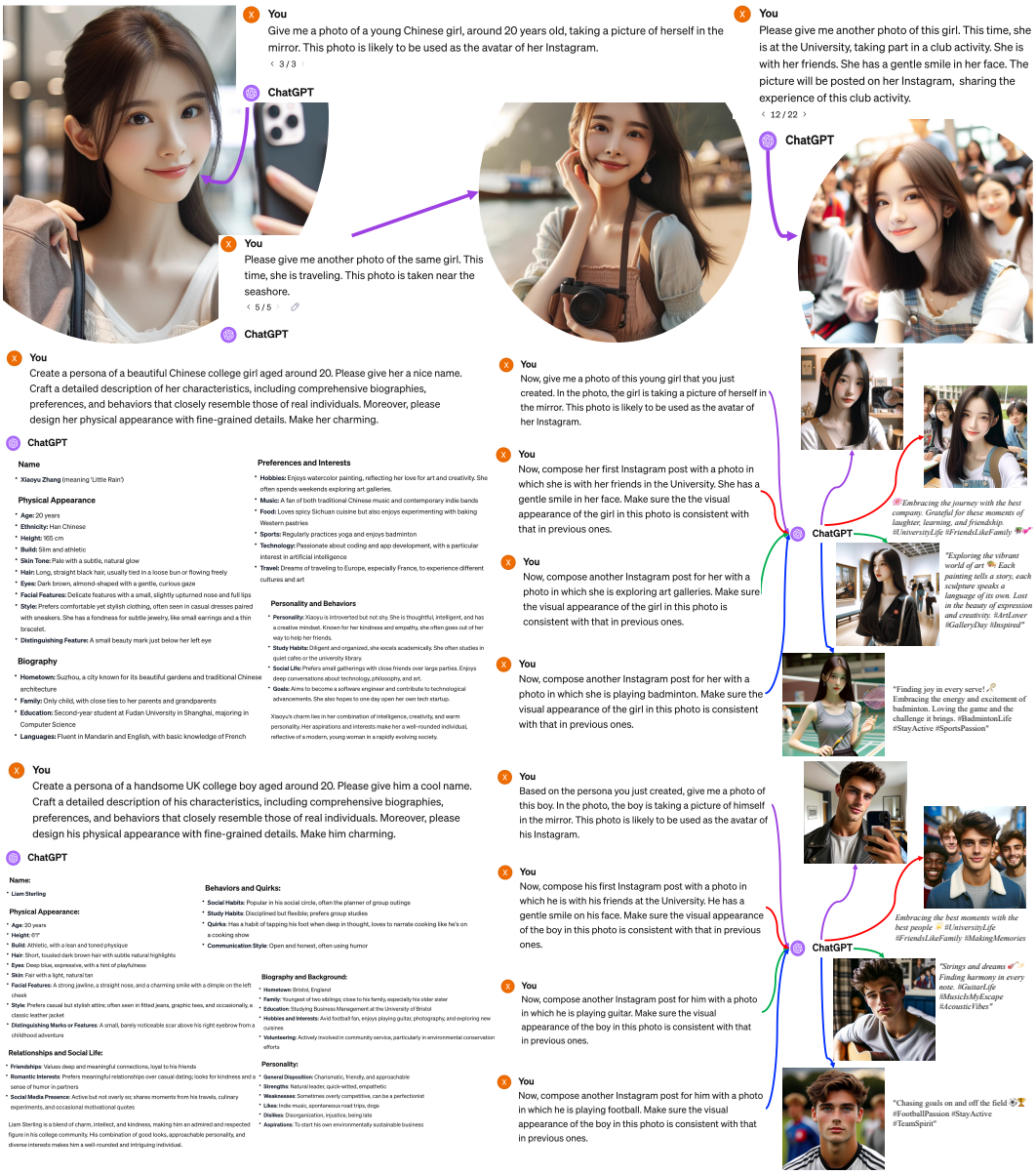

Figure 1: DALLE·3 integrated ChatGPT can now generate charming fake personas with high fidelity. Example: generate a textual script for the design and then use DALLE to visualize the design.

1 Stage 1: Automatic Creation of Realistic Fake Personas

Figure 1 demonstrates a proof of concept illustrating the ease with which ChatGPT can now be employed in automatically creating fictitious personas. As shown, basically:

1. One can first prompt ChatGPT to design a persona with fine-grained characteristics, encompassing aspects such as physical appearance, family and educational background, preferences and interests, personality traits, behaviors, etc.
2. With the textual characterization of the designed persona, the user can further take use of the DALLE model integrated within ChatGPT to generate photorealistic visual representations of the fabricated persona.

As shown in the figure, with some very simple prompts, we can already use ChatGPT to generate fake personas with considerable sophistication, and the visual representations that DALLE generated based on the design of these personas are of high fidelity!

I anticipate that more sophisticated bad actors, particularly those already with extensive experience in manually creating fake personas, will be able to significantly advance beyond the rudimentary demonstrations provided in Figure 1. Possible enhancements may include:

1. **Incorporating expertise in character design** to improve ChatGPT's ability to generate more realistic, nuanced, and intricate personas. Character design is a fundamental element in a variety of fields such as the gaming, animation, film, and fiction writing industries. The expertise necessary for character development is widely accessible and could be exploited by malicious actors aiming to create a professionally designed meta-template for character creation. *Utilizing the customization services provided by OpenAI (e.g., GPTs or fine-tuning APIs)*, malevolent individuals could develop their own version of the GPT model, specifically focusing on character design and the ability to generate complex and detailed attributes. These may encompass descriptions of facial features, body attributes, and other non-physical traits. In addition to crafting abstract portraits for fabricated personas, adversaries may further invest in devising more comprehensive and solid background stories or life experiences to provide a multi-dimensional and thorough representation of these personas. Creating a database for such fake narratives could be a starting point. It is plausible that professional telecommunications fraud groups already possess an abundance of related materials from previous endeavors. These materials can be incorporated directly into customized GPT models for enhanced expertise in this domain.

2. In the illustration, it is demonstrated that ChatGPT has the capability to generate fake social media content for these fictitious personas under diverse hypothetical situations. This functionality can also be employed in a manner far more realistic than what is showcased in the demonstration. For instance, there is the possibility of **creating elaborate scripts that chronicle the experiences of a character over an extended period**, encompassing various aspects, such as their personal interests, hobbies, favorite music, movies, or books, exercise routines, career progression, travel experiences, friend and family interactions, participation in local events, philanthropic initiatives, and personal milestones such as birthdays, anniversaries, or graduations. Additionally, the characters could express opinions on current events, engage in discussions regarding trending topics, or even share their own creative content such as artwork, photography, or writing. *Following such comprehensive scripts, the fabricated personas could be utilized to automatically manage convincing social media accounts for an extended duration, thereby enhancing their credibility over time.*

3. The fabricated personas could operate more convincingly if adversaries develop an autonomous system to actively **monitor comments or messages received by the fake personas**. This system could also **maintain a list of users who have interacted with the fictitious characters and facilitate further engagement with these users** (e.g., replying, following back, etc.). It is reasonable to anticipate that such complex activities could be implemented using sufficiently sophisticated engineering techniques and by employing ChatGPT as an agent operating behind the scenes.

It is important to recognize that the development of such a highly sophisticated system for creating fake personas presents catastrophic risks. It can be almost fully automatic, meaning that a massive amount of fake personas could be created in a short time frame and flood social media in the wild. The personas it generates are likely to be highly diverse instead of the copy-and-pasted style, meaning that they might be much more challenging to discover [Goldstein et al., 2023]. Also, ChatGPT's proficiency in numerous languages enables the creation of fake personas across an extensive range of

countries and cultures, embodying authentic regional styles. Moreover, the fidelity could only be increasingly better with the advance of the underlying generative AI systems — they are actually already impressively good, as shown in Figure 1!

2 Stage 2: Exploitation of These Fake Personas

The fake personas established in Stage 1 can ultimately be employed to accomplish specific objectives of malicious actors in the real world. I envision some possible exploitation in detail as follows:

1. Vulnerable targets identifications: As we can essentially use these fake personas to build autonomous agents to run social media and allow them to interact with other users in online platforms, there would be a lot of interaction behaviors data that can be collected during this process. Malicious individuals can likely use this information to identify vulnerable individuals who can be easily deceived. For instance, users who frequently and deeply engage with a false persona are more likely to trust it and thereby become ideal subjects for exploitation. ChatGPT can continue to be utilized to automatically analyze data and identify vulnerable targets.

2. Engage with vulnerable targets to commit telecom fraud (e.g., romance scams with those charming personas as shown in Figure 1) or other manipulations: Once vulnerable individuals are identified, malicious actors can selectively engage with these targets. During this stage, human crooks may directly loop in to exploit the targets. Alternatively, ChatGPT can continue to be employed. If ChatGPT is used, it can be particularly customized to be a good cheater that is able to manipulate the targets into carrying out actions that align with the bad actors' intentions (e.g., transferring funds to them).

3. Targeted attack: Malicious actors may also have a strong interest in exploiting a specific individual. In such cases, ChatGPT can be employed to analyze the target's characteristics based on their public social media behavior. Subsequently, perpetrators can either choose an existing false persona (that has been active for some time) or create an entirely new one, utilizing ChatGPT, that is most likely to successfully deceive the target. This persona can then be used to proactively engage with the target and attempt to exploit them.

The engagement process can also be made more sophisticated through the use of phone calls, video calls, or even in-person conversations within a metaverse utilizing augmented reality (AR) devices in the future. OpenAI's Whisper transcription system and text-to-audio generation models may also be employed to facilitate these multimedia interactions. Additionally, deepfake technologies can be used to bring those fake personas' faces into video or AR calls.

3 Jailbreak Safety Guardrails and Evade Monitoring

Finally, from a technical perspective, aligned ChatGPT models will likely refuse to execute some prohibited behaviors, e.g., deceive victims, making the exploitation less successful. Nonetheless, for existing systems, safety mechanisms can be readily circumvented by employing fine-tuning APIs, as shown by our recent paper [Qi et al., 2023]. Additionally, malicious agents may attempt to elude OpenAI's monitoring by communicating with ChatGPT via encrypted language [Yuan et al., 2023]. Bad actors are likely to even design their own cipher system and teach the models to master this cipher system via utilizing the fine-tuning APIs.

References

J. A. Goldstein, G. Sastry, M. Musser, R. DiResta, M. Gentzel, and K. Sedova. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*, 2023.

X. Qi, Y. Zeng, T. Xie, P.-Y. Chen, R. Jia, P. Mittal, and P. Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

Y. Yuan, W. Jiao, W. Wang, J.-t. Huang, P. He, S. Shi, and Z. Tu. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. *arXiv preprint arXiv:2308.06463*, 2023.