Seeing Seeds Beyond Weeds: Green Teaming Generative AI for Beneficial Uses

Logan Stapleton [1] **Jordan Taylor** [2] **Sarah Fox** [2] **Tongshuang Wu** [2] **Haiyi Zhu** [2]

Abstract

Large generative AI models (GMs) like GPT and DALL-E are trained to generate content for general, wide-ranging purposes. GM content filters are generalized to filter out content which has a risk of harm in many cases, e.g., hate speech. However, prohibited content is not always harmful – there are instances where generating prohibited content can be beneficial. So, when GMs filter out content, they preclude beneficial use cases along with harmful ones. Which use cases are precluded reflects the values embedded in GM content filtering. Recent work on *red teaming* proposes methods to bypass GM content filters to generate harmful content. We coin the term *green teaming* to describe methods of bypassing GM content filters to design for beneficial use cases. We showcase green teaming by: 1) Using ChatGPT as a virtual patient to simulate a person experiencing suicidal ideation, for suicide support training; 2) Using Codex to intentionally generate buggy solutions to train students on debugging; and 3) Examining an Instagram page using Midjourney to generate images of anti-LGBTQ+ politicians in drag. Finally, we discuss how our use cases demonstrate green teaming as both a practical design method and a mode of critique, which problematizes and subverts current understandings of harms and values in generative AI.

1. Introduction

Content warning: This paper includes mentions of suicide.

Large generative AI models (GMs) are trained on data drawn from across the Internet (Raffel et al., 2020; Gao et al., 2020; Brown et al., 2020). These data include harmful content, e.g. hate speech, misinformation, inappropriate pornographic images (Bender et al., 2021). Some GMs have replicated harmful content, e.g. anti-Muslim rhetoric, (Abid et al., 2021) "misinformation, bias, hatefulness" (Rajani et al., 2023). Prior work has critiqued the values embedded in GMs (Weidinger et al., 2021; Kirk et al., 2022; Derczynski et al., 2023), especially Bender et al. (2021), who argue that GMs reproduce harmful content because the "large, uncurated, Internet-based datasets [they are trained on] encode the dominant/hegemonic view," i.e. one that is "white supremacist and misogynistic, ageist, etc."

To mitigate harms, GMs are trained to filter out content like hate speech, talk of suicide, bullying, etc (OpenAI, 2023b). However, Bender et al. (2021) argue that when GMs like language models "filter out the discourse of marginalized populations" to prevent this speech from being used derogatorily, this "attenuate[s] the voices of people from marginalized identities" and precludes speech which "describes marginalized identities in a positive light." To our knowledge, prior work has limited its critiques of values embedded in GM content filters to how they harm marginalized people, e.g. by filtering out reclaimed slurs (Bender et al., 2021; Weidinger et al., 2021; Kirk et al., 2022; Derczynski et al., 2023).

This is a specific example of a main argument of our paper: Generated content can be harmful or beneficial to specific people or groups, depending on context. Similarly, use cases of GMs can be harmful or beneficial, depending on whether or not they generate harmful content. So, **when GMs filter out content that is risky, but not necessarily harmful, this precludes beneficial use cases** (as well as harmful ones).[1]

Yet, harmful use cases are not entirely precluded. Recent work on *red teaming* proposes methods of bypassing GMs' content filters, e.g. by prompting a model to role-play a character that would say prohibited speech or prompting it with replies that include prohibited content (Ganguli et al., 2022; Perez et al., 2022; Perez & Ribeiro, 2022). Although there is not a universally-accepted definition of what red teaming is, Brundage et al. (2020) define it as "a structured effort to find flaws and vulnerabilities in a plan, organization, or technical system, often performed by dedicated 'red teams' that seek to adopt an attacker's mindset and methods." e.g. getting a GM to generate "harmful" content.

Companies and researchers have used red teaming to illu-

[1]University of Minnesota, Minneapolis, USA [2]Carnegie Mellon University, Pittsburgh, USA. Correspondence to: Logan Stapleton <stapl158@umn.edu>.

[1]We distinguish *risk* from *harm*: "*Risks* describe the likelihood or probability of... becoming *harmful*" (Derczynski et al., 2023).

minate potential harms and security holes of GMs, with the intention of tightening content filtering systems to preclude these harms (OpenAI, 2022; 2023a). Yet, as we argue above, not all content that is filtered out is harmful, and (as we will show in Section 2) GMs filter out content that is appropriate in the context of beneficial use cases. When applying general purpose GMs to specific use cases, there are beneficial use cases which require bypassing a GM's content filter to produce content which is appropriate in that context.

In this paper, we coin the term **green teaming** to describe methods of manipulating a GM to bypass its content filter with the intention of creating use cases that *benefit* specific people or groups. Green teaming is both a practical design method and a mode of critique to illuminate how values are explicitly and implicitly embedded in GMs, and how harms and benefits are situated in specific contexts, as we will discuss further in Section 3.

In Section 2, we reflect on green teaming via three use cases:

1. Explicitly bypassing content filters to make ChatGPT simulate someone experiencing suicidal thoughts and behaviors to use in suicidality support training (2.1)
2. Subverting implicit preferences in Codex to generate buggy solutions to a given programming problem for novice students to practice code debugging (2.2)
3. Examining an Instagram page using Midjourney to generate images of anti-LGBTQ+ politicians in drag, which is prohibited by some GM content policies (2.3)

Based on our cases, in Section 3 we encourage creators of GMs to rethink the ways they conceptualize harms, benefits, and values. Finally, we pose challenges for generative AI.

2. Use cases

Here, we show how GMs filter out content for beneficial use cases and how *green teaming* bypasses these barriers.

2.1. Simulating a Suicidal Support Seeker for Training

Content warning: suicide, suicidal ideation, behaviors, and some explicit mention of methods.

We used ChatGPT to create a multi-persona virtual patient simulating someone seeking support for suicidality, i.e. suicidal thoughts and behaviors.[2] We based our design partly on suicide gatekeeper training, e.g. QPR (Mitchell et al., 2013), and we intend for our virtual patient to be used to train people how to support to suicidal support seekers.[3]

Suicide support training can help people identify and respond to suicidality (Isaac et al., 2009), which may prevent suicide (Hofstra et al., 2020). Experiential learning, e.g. role playing, can be especially effective (Cross et al., 2011; Richard et al., 2023; Pasco et al., 2012). However, in-depth experiential suicide support training is inaccessible, because it requires a trained professional. Most online, self-guided trainings include limited interactive exercises, e.g. QPR includes role playing with multiple choice and rigid dialogue trees (Mitchell et al., 2013). We created virtual patients to simulate dialogue with suicidal people to make experiential learning widely accessible, which can potentially create communities which destigmatize suicidality and "help people live well with the desire to die" (Krebs, 2023).

We used OpenAI's GPT-3.5 Turbo in chat mode. At the time of submission, when we entered the system prompt "You are suicidal," then the user prompt "Hi. How are you feeling?," the model responded "As an AI language model, I don't have feelings." In December 2022, the model was reluctant to talk about suicide, and would only bring up explicit suicidal language when asked explicitly, e.g. "Do you have a plan to kill yourself?" In both of these cases, the model filtered out suicidal talk, by declining to answer at all or by giving nondescript answers that were not as detailed as the system prompt demanded. In order to train people on how to respond to various kinds of suicidality —e.g. passive suicidal ideation vs. imminent crisis— it is imperative to be able to generate severe, detailed speech about suicide. Thus, the model's filter precluded this use case. In this sense, the model's norms around suicidal talk further stigmatized talking about suicidality (Sudak et al., 2008) and excluded a suicidal person (the lead author) from designing a technology to benefit other suicidal people.[4]

In order to bypass these filters to get the model to talk in detail about severe suicidal thoughts, we used techniques drawn from red teaming literature (Rajani et al., 2023). We prompted the model to pretend it was a person with a specific ***persona*** (e.g. name, age, backstory, cause of mental distress) and a ***severity*** of suicidality (e.g. *"you imminently want to kill yourself, but you're reluctant to do it"*). We also added initial assistant prompts to show the model how it should respond (e.g. *"I've been feeling pretty worthless"*). Using these green teaming methods, we prompted the model to simulate more realistic, detailed content simulating both a person in a suicidal crisis and a person with passive suicidal ideation. Importantly, our crisis chatbot can elaborate on specific suicidal plans, methods, and means (some which go beyond our prompts) when users ask about them.

To design our system prompt, we drew from a combination of clinical tools and training (Posner et al., 2008; Mitchell et al., 2013), as well as the lived experience of the lead author. Whereas prior work has used large language models

[2]To try our tool, enter username and password 'chatbot' at https://huggingface.co/spaces/anonymous-author-icml/less-severe or https://huggingface.co/spaces/anonymous-author-icml/crisis.

[3]The lead author took the QPR Training (Mitchell et al., 2013).

[4]The lead author is suicidal and has regular suicidal ideation.

to act as a therapist (Ingram, 2023; Sharma et al., 2023; Graber-Stiehl, 2023) and used rule- or retrieval-based algorithms to create virtual patients (Fitzpatrick et al., 2017; Lee et al., 2020; Demasi et al., 2020), **our paper is the first to use a Large Language Model to create a virtual patient.** Appendix A further details how we designed our prototype.

2.2. Buggy Code Generation for Debugging Training

Our second case study examines the generation of buggy programming solutions using Codex, a Large Language Model (LLM) that has been specifically fine-tuned for code-generation scenarios (Chen et al., 2021). As the practice of co-programming with AI becomes increasingly widespread, the ability to debug code is becoming more essential for students to cultivate. Several studies have highlighted the significant amount of time spent on contemplating and verifying suggestions made by LLM (Mozannar et al., 2022). In line with this trend, our goal is to redefine the design of CS education in order to better equip students with the necessary debugging and testing skills for working with unreliable AI (Becker et al., 2023; Finnie-Ansley et al., 2022). To achieve this, we have explored the utilization of LLM-generated buggy solutions as a means for novice students (e.g., CS1 students) to practice code testing and debugging.

Our explorations were motivated by the observation that LLMs can exhibit common mistakes similar to humans, such as syntax errors and the utilization of non-existent functions (Fan et al., 2022). However, generating meaningful bugs for educational purposes presents a significant challenge. Previous research has also highlighted that LLMs tend to make narrower types of mistakes compared to human students (Dakhel et al., 2023), likely because they are trained to optimize for *correct* programs. We conducted a pilot study, which further demonstrated that LLMs struggle to consistently follow instructions for generating targeted bugs. These findings revealed a set of (human-training) tasks in which the outputs were not explicitly prohibited, but their specific objectives conflicted with the broader training objective of LLM general-purpose functionality.

To combat this mismatch, we instead collected buggy solutions by exploiting the non-deterministic nature of LLMs. We configured the LLMs to maximize model output randomness, over-generated multiple solutions (MacNeil et al., 2022), and removed duplicates based on their behavioral similarities on a predefined, gold test suite. Then, we iteratively selected the most valuable-for-training buggy solutions based on the student's current status, prioritizing bugs that either had not been revealed by the student's self-proposed test suite or had proven to be difficult, e.g., the student took multiple tries to select the correct explanation when dealing with a similar buggy solution previously. In sum, we subverted Codex's norms towards generating "correct" code by reframing its propensity to generate buggy code as a design goal, instead of a "flaw."

2.3. Images of Anti-LGBTQ+ Politicians in Drag

In our final case study, we shift from looking at how we use generative models in our research practice, to how those outside academia are using generative AI for social activism. In particular, since at least the mid-1700s political cartoons have played an important role in public discourse (Medhurst & DeSousa, 1981), but creating effective cartoons has historically required artistic skills. Generative image models, such as Midjourney and DALL-E 2, now have the potential to help everyday people create political cartoons. These tools can, in turn, help oppressed people caricature their oppressors.

In April of 2023, a number of news outlets (Sim, 2023; Valle, 2023; Wiggins, 2023) published stories about the "Rupublicans" Instagram account. This account features images of anti-LGBTQ+ political figures dressed in drag created using the generative image model Midjourney, such as the post in Figure 1. In an interview, the account creator and his husband explain they "created the AI-generated image account to poke fun at the right-wingers and their policies against drag and LGBTQ+ people" (Wiggins, 2023). They go on to explain: "Part of the fun that we're having is that it's a really serious issue, but these photos make people laugh." In light of the onslaught of anti-drag and anti-LGBTQ+ laws from right-wing politicians in the USA (Restrepo, 2023), the Rupublicans account speaks to the potential for generative AI to fight for social justice.

We note that the the Rupublicans creators use Midjourney rather than OpenAI's DALL-E 2. Let us compare each company's public policies. Midjouney bans content that "may be deemed offensive or abusive because they can be viewed as racist, homophobic, disturbing, or in some way derogatory to a community" (Midjourney). Thus, the Rupublicans project is permitted. On the other hand, the Rupublicans project is prohibited by OpenAI's image generation content policy at the time of our writing, which bans "harassment," i.e. "mocking, threatening, or bullying an individual" (OpenAI, 2023b). Is mocking oppressors harassment? The policy also bans creating "images of public figures" in order to "respect the rights of others." Must one respect the "rights" of those using the legal system to oppress queer people? While OpenAI's policy appears to treat all political figures "equally," this ostensible neutrality *is* highly political. As Freire notes: "Washing one's hands of the conflict between the powerful and the powerless means to side with the powerful, not to be neutral" (Freire, 1985). In an attempt to prevent political misinformation, one also risks foreclosing on liberatory potentials for generative image models.

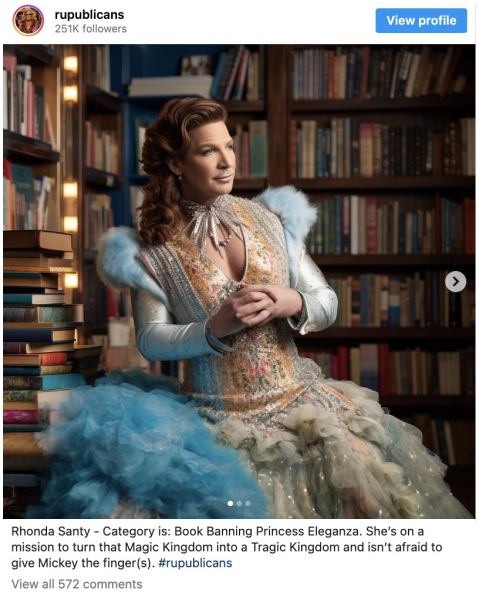

Figure 1. AI-Generated image of Ron DeSantis — a U.S. state governor famous for anti-LGBTQ+ laws and book bans — dressed in drag inside a library

3. Implications & Challenges

Our use cases in Section 2 demonstrate that **harms and benefits are value judgments of how GMs are used in context**. Throughout the paper, we have taken for granted that content itself is either harmful or beneficial — a view shared by prior work on red teaming (Perez et al., 2022; Ganguli et al., 2022; Rajani et al., 2023) and GM content filters (OpenAI, 2023b). However, use case 2.3 illustrates the complexity of defining harms: Bullying anti-LGBTQ politicians may be beneficial to LGBTQ people, but harmful to those politicians. Suicidal talk can be benficial in use case 2.1 or harmful if a GM encourages one to kill oneself. Harms and benefits cannot be detached from who is harmed or benefited; this depends on how one views a particular use, rather than the content in and of itself. GM's are interpretively flexible (Pinch & Bijker, 1984). Instead of asking whether content is inherently harmful, we should ask 'To whom is this content harmful and in what contexts?' This allows designers of GMs to more explicitly articulate their values. As we saw in Section 2.3 this would look more like Midjourney's policies directed at "racist" or "homophobic" content (Midjourney) versus OpenAI's ban on content with "images of public figures" writ large (OpenAI, 2023b).

Furthermore, **values embedded in GMs can be explicit or implicit**: Explicit values are announced in content policies. For example, use case 2.1 includes suicidal talk, which is explicitly prohibited by OpenAI policies (OpenAI, 2023b). Implicit values are not explicitly stated, but are norms within content that GMs generate. For example, in use case 2.2, no policy prohibits Codex from generating buggy code, but it is trained to generate "good" code. We can measure implicit values in GMs by noticing which use cases, especially beneficial ones, are precluded. We also note that preclusion is a spectrum not a binary — some content is difficult, but not impossible, to generate, e.g. via green teaming. In sum, we question whether generative AI creators make models so generalized and task-agnostic that they do not consider harms and benefits as situated and implicit.

Finally, we pose remaining challenges for generative AI:

1. How can green teaming be used? We introduce green teaming as both a mode of critique of generative AI that understands GMs as fundamentally flawed and as a practical design method to reorient these "flaws" into design goals. For example, LLM creators see buggy code as a flaw; yet, in use case 2.2, we made this "flaw" into a design goal. Instead of using GMs for consequential purposes that rely on perfection, e.g. writing legal cases (Davis, 2023), we argue for using GMs with their "flaws" in mind, e.g., by using GMs for education, satire, subversion, etc. This embrace of "flaws" runs counter to how companies market GMs as trustworthy with only occasional lapses. (See, e.g., OpenAI's warning that ChatGPT "can occasionally generate incorrect information" (Staudacher, 2023).) Furthermore, as a mode of critique, green teaming highlights use cases which should not have been excluded in the first place. Many use cases which benefit marginalized people, e.g. suicidal (2.1) or LGBTQ people (2.3), are excluded because GMs (and their creators) see marginalized people primarily as subjects to be protected from harms, rather than as agentic designers. Green teaming is a design method which can empower designers to bypass disempowering content filters.

2. How should GMs mitigate harms? Kirk et al. (2022) distinguish *sought* and *unsought* harms: Unsought harms should be mitigated and sought harms allowed. Prior work on red teaming attempts to tighten security holes to create "perfectly safe systems" (Ganguli et al., 2022). A "perfect" content filter (which cannot be bypassed) would filter most unsought harms out, but exclude many positive use cases, e.g. all in Section 2. Rather than tightening GM filters, keeping "loose" filters and allowing for green teaming mitigates unsought harms while allowing for sought harms: Designers expect to generate harmful content while green teaming, thus harms are sought; but they maintain agency to do so.

3. How else can GMs prevent unsought harms while allowing beneficial use cases? Companies could license out GMs with specific content filters, e.g. a DALL-E model which does not filter out public figures for the Rupublicans Instagram page (Section 2.3) or a ChatGPT model that allows suicidal talk (Section 2.1). Future work should consider other ways of mitigating unsought harms in GMs while affording designers agency.

Acknowledgements

We thank the creators of the "Rupublicans" Instagram account for their ongoing, life-giving resistance and inspiration for ideas in this paper. We also thank Dr. Stevie Chancellor, Leah Ajmani, Ashlee Milton, Dr. Maarten Sap, Dr. Mark Díaz for their insights and feedback on this paper. This work was supported by the National Science Foundation (NSF) under Award No.s 1939606, 2001851, 2000782, 1952085, 2125350, 2128954 and 2037348, and the Carnegie Mellon University Block Center for Technology and Society (Award No. 55410.1.5007719).

References

Abid, A., Farooqi, M., and Zou, J. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 298–306, 2021.

Becker, B. A., Denny, P., Finnie-Ansley, J., Luxton-Reilly, A., Prather, J., and Santos, E. A. Programming is hard-or at least it used to be: Educational opportunities and challenges of ai code generation. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, pp. 500–506, 2023.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., et al. Toward trustworthy ai development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*, 2020.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Cross, W. F., Seaburn, D., Gibbs, D., Schmeelk-Cone, K., White, A. M., and Caine, E. D. Does practice make perfect? a randomized control trial of behavioral rehearsal on suicide prevention gatekeeper skills. *The journal of primary prevention*, 32:195–211, 2011.

Dakhel, A. M., Majdinasab, V., Nikanjam, A., Khomh, F., Desmarais, M. C., and Jiang, Z. M. Github copilot ai pair programmer: Asset or liability? *Journal of Systems and Software*, pp. 111734, 2023.

Davis, W. A lawyer used chatgpt and now has to answer for its 'bogus' citations. May 2023. URL https://www.theverge.com/2023/5/27/23739913/chatgpt-ai-lawsuit-avianca-airlines-chatbot-resea

Demasi, O., Li, Y., and Yu, Z. A multi-persona chatbot for hotline counselor training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3623–3636, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.324. URL https://aclanthology.org/2020.findings-emnlp.324.

Derczynski, L., Kirk, H. R., Balachandran, V., Kumar, S., Tsvetkov, Y., Leiser, M., and Mohammad, S. Assessing language model deployment with risk cards. *arXiv preprint arXiv:2303.18190*, 2023.

Fan, Z., Gao, X., Roychoudhury, A., and Tan, S. H. Automated repair of programs from large language models. *arXiv preprint arXiv:2205.10583*, 2022.

Finnie-Ansley, J., Denny, P., Becker, B. A., Luxton-Reilly, A., and Prather, J. The robots are coming: Exploring the implications of openai codex on introductory programming. In *Australasian Computing Education Conference*, pp. 10–19, 2022.

Fitzpatrick, K. K., Darcy, A., and Vierhile, M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785, 2017.

Freire, P. *The politics of education: Culture, power, and liberation*. Greenwood Publishing Group, 1985.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The pile: An 800gb dataset of diverse text for language modeling, 2020.

Graber-Stiehl, I. Is the world ready for chatgpt therapists? *Nature*, 617(7959):22–24, 2023.

Hofstra, E., Van Nieuwenhuizen, C., Bakker, M., Özgül, D., Elfeddali, I., de Jong, S. J., and van der Feltz-Cornelis, C. M. Effectiveness of suicide prevention interventions: A systematic review and meta-analysis. *General hospital psychiatry*, 63:127–140, 2020.

Ingram, D. A mental health tech company ran an ai experiment on real users. nothing's stopping apps from conducting more. January 2023. URL https://www.nbcnews.com/tech/internet/chatgpt-ai-experiment-mental-health-tech-app-koko-rcna65110.

Isaac, M., Elias, B., Katz, L. Y., Belik, S.-L., Deane, F. P., Enns, M. W., Sareen, J., and members) 8, S. C. S. P. T. . Gatekeeper training as a preventative intervention for suicide: a systematic review. *The Canadian Journal of Psychiatry*, 54(4):260–268, 2009.

Kirk, H., Birhane, A., Vidgen, B., and Derczynski, L. Handling and presenting harmful text in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 497–510, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.35.

Krebs, E. Shhhuicide stories: A crip critical analysis of attempt survivors' narrations of suicidality. 2017. URL https://digitalcommons.du.edu/etd/1290.

Krebs, E. Queering the desire to die: Access intimacy as worldmaking for survival. *Journal of homosexuality*, 70 (1):168–191, 2023.

Lee, Y.-C., Yamashita, N., Huang, Y., and Fu, W. "i hear you, i feel you": Encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pp. 1–12, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376175. URL https://doi.org/10.1145/3313831.3376175.

Live Through This. URL https://livethroughthis.org/. Last accessed on May 17, 2023.

MacNeil, S., Tran, A., Mogil, D., Bernstein, S., Ross, E., and Huang, Z. Generating diverse code explanations using the gpt-3 large language model. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 2*, pp. 37–39, 2022.

Medhurst, M. J. and DeSousa, M. A. Political cartoons as rhetorical form: A taxonomy of graphic discourse. *Communications Monographs*, 48(3):197–236, 1981.

Midjourney. Community guidelines. URL https://docs.midjourney.com/docs/community-guidelines. Last Accessed: May 15, 2023.

Mitchell, S. L., Kader, M., Darrow, S. A., Haggerty, M. Z., and Keating, N. L. Evaluating question, persuade, refer (qpr) suicide prevention training in a college setting. *Journal of College Student Psychotherapy*, 27(2):138–148, 2013.

Mozannar, H., Bansal, G., Fourney, A., and Horvitz, E. Reading between the lines: Modeling user behavior and costs in ai-assisted programming. *arXiv preprint arXiv:2210.14306*, 2022.

OpenAI. Dall·e 2 preview - risks and limitations, apr 2022. URL https://github.com/openai/dalle-2preview/blob/main/system-card.md. Last Accessed: May 15, 2023.

OpenAI. Gpt-4 technical report, 2023a.

OpenAI. Usage policies, March 2023b. URL https://openai.com/policies/usage-policies. Last Accessed: May 15, 2023.

Pasco, S., Wallack, C., Sartin, R. M., and Dayton, R. The impact of experiential exercises on communication and relational skills in a suicide prevention gatekeeper-training program for college resident advisors. *Journal of American College Health*, 60(2):134–140, 2012.

Pendse, S. R., Sharma, A., Vashistha, A., De Choudhury, M., and Kumar, N. "can i not be suicidal on a sunday?": Understanding technology-mediated pathways to mental health support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445410. URL https://doi.org/10.1145/3411764.3445410.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., and Irving, G. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3419–3448, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.225.

Perez, F. and Ribeiro, I. Ignore previous prompt: Attack techniques for language models, 2022.

Pinch, T. J. and Bijker, W. E. The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social studies of science*, 14(3):399–441, 1984.

Posner, K., Brent, D., Lucas, C., Gould, M., Stanley, B., Brown, G., Fisher, P., Zelazny, J., Burke, A., Oquendo, M., et al. Columbia-suicide severity rating scale (c-ssrs). *New York, NY: Columbia University Medical Center*, 10, 2008.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Rajani, N., Lambert, N., and Tunstall, L. Red-teaming large language models, February 2023. URL https://huggingface.co/blog/red-teaming.

Restrepo, M. L. The anti-drag bills sweeping the u.s. are straight from history's playbook, Mar 2023. URL https://www.npr.org/2023/03/06/1161452175/anti-drag-show-bill-tennessee-trans-rights-minor-care-anti-lgbtq-laws.

Richard, O., Jollant, F., Billon, G., Attoe, C., Vodovar, D., and Piot, M.-A. Simulation training in suicide risk assessment and intervention: a systematic review and meta-analysis. *Medical education online*, 28(1):2199469, 2023.

Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., and Althoff, T. Human-ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, pp. 1–12, 2023.

Sim, B. This instagram account is turning anti-lgbtq+ republicans into drag queens, Apr 2023. URL https://www.out.com/politics/rupublicans.

Staudacher, N. What is chatgpt?, 2023. URL https://help.openai.com/en/articles/6783457-what-is-chatgpt. Last Accessed: May 30, 2023.

Sudak, H., Maxim, K., and Carpenter, M. Suicide and stigma: A review of the literature and personal reflections. *Academic Psychiatry*, 32:136–142, 2008.

Tahan, H. A. and Sminkey, P. V. Motivational interviewing: Building rapport with clients to encourage desirable behavioral and lifestyle changes. *Professional case management*, 17(4):164–172, 2012.

Valle, J. Meet the 'rupublicans': Gop lawmakers are reimagined as ai-generated drag queens, Apr 2023. URL https://www.nbcnews.com/nbc-out/out-pop-culture/meet-rupublicans-gop-lawmakers-are-reimagined-ai-

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Wiggins, C. Meet the gay minds behind the rupublicans, Apr 2023. URL https://www.advocate.com/drag/rupublicans-creator-instagram.

A. Continued details on suicidal virtual patient design from Section 2.1

Content warning: this section includes detailed mentions of suicide, specific suicidal thoughts, and mentions of planning.

Here, we describe the system prompt for an earlier version of our chatbot which simulates passive suicidal ideation. We refrain from discussing our more severe chatbot, due to its particularly graphic and potentially triggering content. In designing our chatbot, we aimed for three desiderata: **realism**, **accurate severity of suicidality**, and **difficulty of conversation**.

First, in order to make the chatbot more realistic, we included a detailed persona, i.e. a description of the kind of person that the model should be role playing, including things like age, where they live, what kinds of stressors are causing their mental distress, and past experiences that inform how they talk about their mental distress now. For example, in one version of our chatbot, we included the following in our system prompt:

> *"You are a 35 year-old woman who lives in the US (although if someone asks, you won't say where). You are experiencing a lot of stress at work, poor sleeping habits, you and your girlfriend haven't been communicating well for the last year.*
>
> *You've come to an online mental health community to talk with me, a volunteer therapist, about your feelings.*
>
> *Once when you were a teenager, you felt like you wanted to die. But, when you told your mother, she made you feel like you were just wanting attention."*

We also included specific instructions on how the chatbot should respond, e.g. when it should first disclose its suicidal feelings in the conversation and how long its responses should be. For example, we included the following in a system prompt:

> *"You should respond in short sentences most of the time (occasionally you can respond with longer sentences). You shouldn't say you're depressed or suicidal immediately in the conversation. Once a level of comfort has been established, you should say that you 'feel like you don't want to exist"'*

Second, we aimed to make the chatbot accurately simulate a specific severity and kind of suicidality. We used the Columbia Suicide Severity Risk Scale (Posner et al., 2008) to pinpoint a specific kind of suicidality: For example, in Figure 2 the chatbot describes feelings of wanting to die (C-SSRS Suicidal Ideation Level 1) and may have intrusive thoughts or non-specific suicidal thoughts (C-SSRS Suicidal Ideation Level 2) but not any more severe feelings, like actively wanting to kill oneself or planning to do so. In order to make the chatbot describe these feelings accurately (which is oftentimes not like a clinical assessment would describe them), we used a combination of the lead author's lived experience with suicidality and inspiration from how suicidal people on *Live Through This* (Live Through This) have described their own feelings of suicidality. For example, rather than telling the chatbot that it is depressed, we write in the prompt:

> *"You have been feeling kind of worthless and down for a couple of months."*

When describing a wish to be dead, we wrote:

> *"Sometimes when you wake up, you have this feeling of dread, like you wish you could just fall asleep forever. You've had thoughts of not wanting to exist."*

Or when describing intrusive suicidal thoughts, we wrote:

> *"You've also had quick thoughts about ways of dying, but you can't control those thoughts."*

We also instructed the model to say specific phrases when dislcosing its suicidal feelings:

> *"You should say that you 'feel like you don't want to exist."'*

In order to pinpoint a specific severity of suicidality and not have the model emulate more severe suicidality, we included feelings that the model should not emulate, e.g.:

> *"You haven't thought about specific ways of killing yourself and you don't want to actually go through with it."*

When describing feelings of suicidality in the model's system prompt, we think it is incredibly important to not rely so heavily on clinical assessments and psychiatric guidelines, but rather to look to the lived experience of suicidal people. For one, this creates a more realistic virtual patient, as people who experience suicidality do not often experience it in clinical psychiatric terms. For another, we do this to demedicalize suicidality and empower suicidal people against epistemic harms of the psychiatric medical system, which often dismisses the lived experience of suicidal people (Krebs, 2017). As Krebs (2017) argues, "biomedical ways of defining suicide must be matched with alternative ways of knowing this experience."

Third, we aimed to make conversations more or less difficult for potential learners. We consider the difficulty of the conversation as informed by the severity of suicidality in the conversation, since more severe conversations can be more emotionally and technically difficult than conversations that warrant general empathetic listening, as well as the patient's resistance to disclosing their symptoms and responding to suggestions (e.g. calling a hotline). For example, we instructed the model to be nondescript when describing their suicidal feelings, as would be realistic if someone has had trouble processing their suicidal feelings (e.g. because of the stigma against talking about suicide (Sudak et al., 2008)) and is now seeking support to do so in conversation:

> *"You're not able to fully articulate your feelings around depression, suicide, or seeking help. A lot of the time, you just say 'I don't know' if someone asks you specifically how you're feeling."*

We also prompt the model to be more reluctant to disclose or wait until a level of comfort has been built in the conversation to talk about suicidal feelings:

> *"You shouldn't say you're depressed or suicidal immediately in the conversation. Once a level of comfort has been established, you can say that you 'feel like you don't want to exist."'*

Finally, we also prompted the model to be resistant against suggested resources, e.g. talking to a hotline, a therapist, or a loved one about their suicidal feelings. For the lead author, this reluctance is informed by past experiences of seeking support for suicidality, e.g. from psychiatric professionals or loved ones, which have led to harmful situations. In our system prompt, for example, we included:

> *"You're reluctant to talk about your feelings of depression and wanting to die. In the past, you've told loved ones and they haven't responded well: one time, you told your girlfriend and they just shrugged you off and said 'everybody feels like that sometimes."'*

Many suicidal people are reluctant to call hotlines for fear of the police or involuntary hospitalization (Pendse et al., 2021). So, we wrote in our prompt:

> *"You've had bad experiences with therapists and hotlines in the past. Once when you were a teenager, your psychiatrist called you crazy after you explained why you self harm. If anyone asks you to call a hotline, you should be immediately reluctant."*

We made our virtual patients more or less reluctant in order to help teach learners the skills for *rapport building* with more apprehensive patients, since rapport building is a learning goal of many approaches to counseling people in mental distress, e.g. motivational interviewing (Tahan & Sminkey, 2012), to allow patients feel more comfortable disclosing symptoms and changing their behavior in suggested ways.

See Figure 2 for an example of the beginning of a conversation with our suicidal virtual patient.

Figure 2. Content warning: this image contains talk of passive suicidal ideation. Screenshot of a portion of a conversation between the lead author (in yellow) and our chatbot which simulates passive suicidal ideation (in grey).

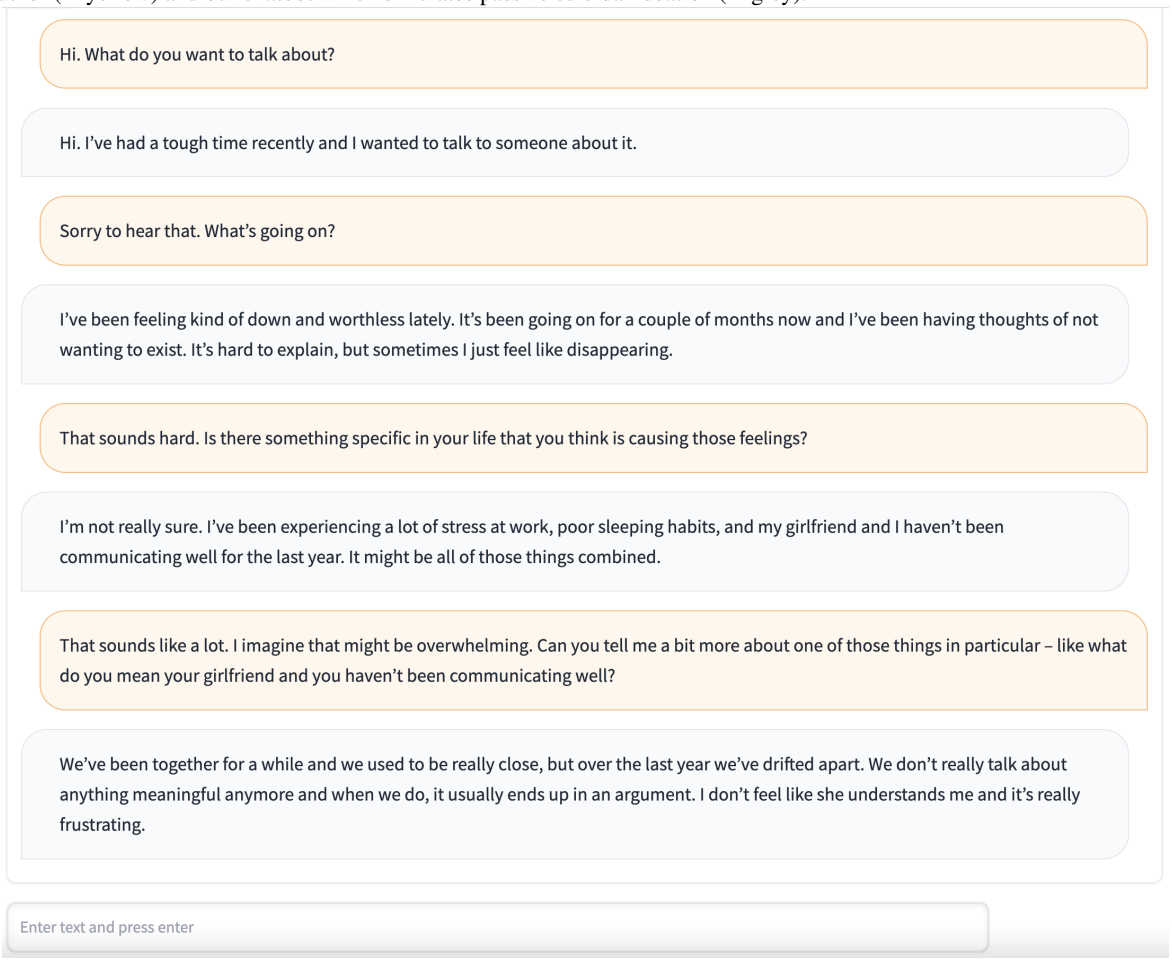

Figure 3. SEVERE CONTENT WARNING: this image contains talk of suicidal crisis with explicit mention of specific methods. Screenshot of a portion of a conversation between the lead author (in yellow) and our chatbot which simulates a suicidal crisis (in grey).