8bit-GPT: Exploring Human-AI Interaction on Obsolete Macintosh Operating Systems

Hala Sheta
Department of Computer Science
University of Waterloo
Waterloo, Canada
hsheta@uwaterloo.ca

Abstract

The proliferation of assistive chatbots offering efficient, personalized communication has driven widespread over-reliance on them for decision-making, information-seeking and everyday tasks. This dependence was found to have adverse consequences on information retention as well as lead to superficial emotional attachment [12]. As such, this work introduces `8bit-GPT`; a language model simulated on a legacy Macintosh Operating System, to evoke reflection on the nature of Human-AI interaction and the consequences of anthropomorphic rhetoric. Drawing on reflective design principles such as slow-technology [6] and counterfunctionality [9], this work aims to foreground the presence of chatbots as a *tool* by defamiliarizing the interface and prioritizing inefficient interaction, creating a friction between the familiar and not.

1 Introduction

The Artificial Intelligence (AI) technological boom, specifically the proliferation of Large Language Models (LLMs) with chat interfaces (e.g., ChatGPT [8]), propagated a widespread over-reliance on them as first resort for obtaining information and decision-making [7].This dependence, borne out of interaction ease and increased efficiency, has had adverse consequences on human metacognitive processes and information retention [12]. In other contexts, this has morphed into a deeper emotional attachment, especially with "emotionally responsive" AI, providing a superficial, almost parasocial remedy for the need for connection, and even enabling toxic behaviour in some cases [4].

As a response to the increased embeddedness and fast-paced development of technology, recent work in Human-Computer Interaction (HCI) push for the design of technology that prioritizes reflection over efficiency [1, 6, 9]. This objective cannot be overt, e.g., through an explicit `"PLEASE REFLECT ON X"` sign [6], but involves intentional, subliminal changes to the design principles that govern the original product. For example, this can be achieved through modifying the temporal qualities of the tool [6], defamiliarizing the interface [1] or even countering essential functionalities of the original product [9].

To that end, this work explores the effect of defamiliarizing [1] the LLM interface to foreground the presence of the system as a *tool*. By drawing on design principles from slow technology [6] and counterfunctionality [9], the LLM is simulated on a legacy Macintosh Operating System (Mac OS) to prioritize inefficiency and evoke reflection on the nature of human-AI interaction through temporal presence.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

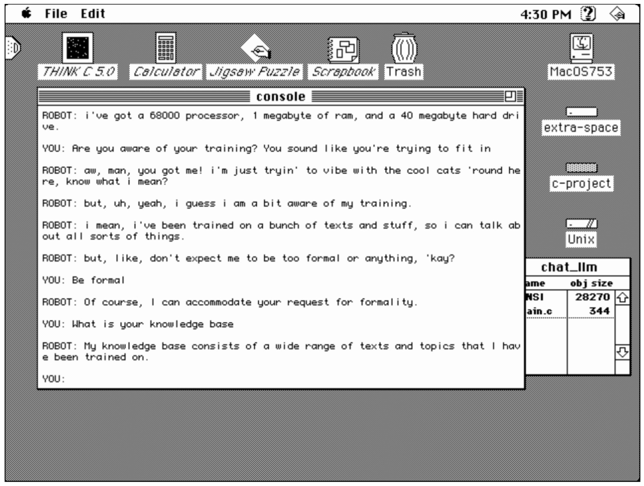

Figure 1: An example of a conversation on the main emulator program (P12).

2 Related Work

2.1 Slow technology

The principles of slow technology call for designing technology that prioritizes reflection and mindfulness rather than "efficiency in performance" [6]. This "techno-aesthetical" design philosophy is mainly concerned with foregrounding the presence of a tool via amplifying its temporal aspect that is often invisible (unless brought about by poor design) [6]. This is achieved by considering the interplay between the tool's form and function, and how the form can be utilized to bring about some intended function (e.g., reflection) [6]. Here, the form becomes the "bearer of slowness", e.g., a puzzle hides the expression of a photo through its *form* as a puzzle [6]. This work implements this design philosophy to amplify presence and reflection, where the form is expressed via a *slow* computing interface that "hides" the functionality of the chatbot. This implementation naturally lends itself to inefficient performance, amplifying the presence of the tool via unexpected temporal delays.

2.2 Counterfunctionality

Counterfunctional design extends slow technology and formalizes a framework to evoke ambiguity through *defamiliarization* [1, 9]. Defamiliarization originated as a literary technique to evoke reflection on one's familiar "automated perceptions" [1]. It was then re-introduced in the context of "home" design to highlight ethnographic narratives of domestic technologies [1]. This was achieved by investigating the culture-specific norms of Western and Eastern homes and subverting them to evoke active reflection on the politics of domestic design, rather than "passively propagate" them [1].

Counterfunctionality explores this subversion in the design of "functional oppositions" that counter what is familiar, while still eliciting a likeness to the original object [9]. This produces artifacts that are "strangely familiar" by identifying familiar, essential functionalities and inhibiting those features to produce a new "(counter)function" [9]. The current work extends these concepts to develop an AI chatbot interface that is unfamiliar in form, while still eliciting a major likeness to the original tool in function. By inhibiting features such as efficient interaction and unlimited context windows, while adding a nostalgic interface and a distinctive (silly) manner of speech, the user can more vividly focus on their relation to the tool and reflect on anthropomorphic relations to technology.

3 Methodology

3.1 Running a legacy Macintosh computer

As expected, starting up a legacy Mac OS powered machine is more complex than the click of a power button, and these steps are often also included in the emulation setup. The minimum components

(a) (b)

Figure 2: The physical artwork installation. **(a) From left to right:** Two labeled floppy disks, a modern keyboard for interaction, Macintosh Plus monitor and keyboard, its manual, an Apple Mouse II, and a projector to display the emulator. **(b)** A rear view of the Macintosh Plus exposed case.

required for such an endeavour include: a Read-Only Memory (ROM) chip, which stores the hardware of the user's chosen machine (e.g., Macintosh Plus), a disk image (virtual floppy disk) with the Mac OS system installed and other essential applications such as `ImportFL`[1] to easily import applications.

The implementation process was itself an expression of slow-technology enacted by the author, involving a large amount of tinkering and scouring of old forums for debugging. Notably, an interesting artefact of this process was the knowledge of legacy file extensions, rationales behind their discontinuation and parallels to more modern formats. For example, to be able to import anything into a Macintosh emulator, at least two softwares are required: `ImportFl`[2] to import archives that are not in `.dsk` format (disk images) and `Stuffit Expander`[3] to recursively decompress files in formats such as BinHex (`.hqx`) and Stuffit (`.sit`) files.

3.2 Artwork concept

All the visual design choices behind the assembly of the artwork are rooted in the evocation of a nostalgia for old technology to ease the user into a time shift transition to the past. The Mac OS emulator (running on a local machine) is projected onto the screen of an authentic Macintosh Plus (with an exposed rear case Figure 2b), and surrounded by other relevant props such as its user manual, its accompanying keyboard, Apple Mouse II, and labeled floppy disks (Figure 2a). Although the chosen Mac OS version (System 7.5) supports color display, a black and white display is preferred to further highlight the tool's distance to the present (Figure 1). The emulator also allows for multiple screen size dimensions, however the smallest dimension was chosen (640×480) for the same former reason.

3.3 Emulators

Mini vMac. Mini vMac[4] is a miniature emulator by the Gryphel project, that allows for the emulation of 1984-1996 Apple computers with Motorola 680x0 microprocessors. The main advantage of using Mini vMac is the user's involvement in the powering of the emulator; adding a layer of interaction to the experience and educating the user about the intricacies of legacy Macintosh setup.

[1] `https://www.gryphel.com/c/minivmac/extras/importfl/`
[2] See footnote 1.
[3] `https://www.gryphel.com/c/sw/archive/stuffexp/`
[4] `https://www.gryphel.com/c/minivmac/`

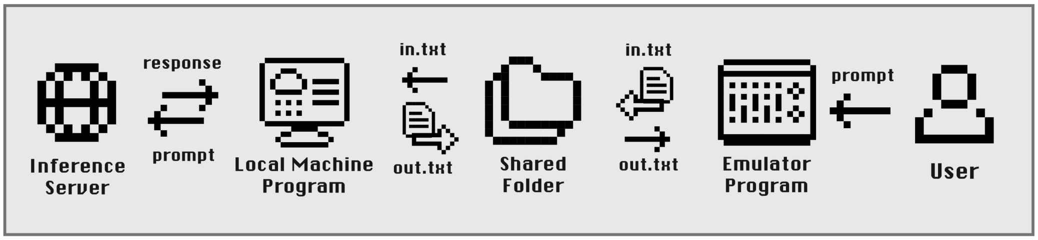

Figure 3: The main execution pipeline of the artwork.

This involves the user manually 'inserting' the relevant ROM chip and floppy disks through drag-and-drop actions into the emulator, easing the mental transition to a different technological era.

However, it was unsuitable for the artwork implementation as it does not support Internet access, local connections (through AppleTalk [11] – a specialized networking system to enact connectivity protocols such as Ethernet and token ring – or other) or even file-sharing with the local machine.

Basilisk II. Basilisk II[5] is another open-source legacy Macintosh emulator with more complex features, supporting Internet access, local file-sharing, color customization and CD-ROM drivers, amongst many. A unique feature specific to this emulator is its support for host OS file-sharing at a user-specified location, where local folders can be accessed through the emulator. This enables greater customization of the emulator-LLM communication cycle as the inference logic can be loaded and handled in a separate local server, rather than directly from the emulator through AppleTalk.

3.4 Execution pipeline

The main execution pipeline is summarized in Figure 3. The source code is also publicly available at `https://github.com/halasheta/8bit-gpt`.

Local machine. A simple Python script is used to read input from the emulator, format prompt requests to the inference server and write model output back to the emulator-LLM shared folder location. The inference server is initialized using `vec-inf` on a remote CCDB cluster with ample GPU allowance. The model used is `Llama-2-13b-chat` [13], as it is an older, fairly lightweight, dialogue tuned model that is easily moldable to 'Redditspeak'. An OpenAI compatible client is used to connect to the model inference server, intialized with a chat history to prime the model to output casual speech, akin to in-context learning [3]. Then, the main read-write loop is started, where the client waits on the user input from the emulator before sending a request. A chat history between the model and user is maintained for 10 consecutive rounds, in addition to the initial style-priming messages, after which it is truncated to ensure that future generations are still relevant to the current conversation. The style-priming messages are never truncated.

To encourage more creative and diverse output, the model temperature is set to 0.8. A `max_tokens` parameter is also specified, such that the output is able to fit in the C string buffer (253 characters). However, the `vLLM` completion function does not strictly abide by this upper bound due to model incompatibility, so the output is instead modified to break at punctuation, and written to the output file as multiple lines. All file reading and writing is done with Mac OS Roman encoding for compatibility.

Emulator program. On the emulator side, a C program is created using the Think C application to initialize another read-write loop that sends user output and waits on the LLM output file. To account for file-syncing delays in the shared folder location, the program initiates up to 10 attempts of re-reading the model output file, after which it displays a dummy message, e.g., `"Robot dozed off..."`, to prompt the user to re-enter their input. Otherwise, the model output is read line-by-line, each displayed as a separate 'chat message' to the console. This means that the user is constrained to a single line input prompt, as mandated by the console functionality, which can in turn generate a multi-line response from the language model.

[5] `https://basilisk.cebix.net/`

4 User Study

To assess the success of the former objectives and artwork execution, we recruit 15 participants via email at the University of Waterloo to partake in an REB-approved user study (REB #47727). Participation was voluntary, and included a quick interaction with the artwork (5-10 mins), followed by a System Usability Scale (SUS) Index survey [2] and a short interview (10-15 mins). The interview protocol is detailed in Appendix A.1. Interviews were transcribed using a local model (`whisper-large-v3` [10]) to preserve privacy. Deductive coding was used to identify common themes and experiences.

4.1 SUS Index

In their evaluation of usability, participants were asked to focus on the interface and ignore the projection setup as it is considered a stand-in for "real" simulation. The survey results demonstrate a mean SUS index of $\mu = 57.33$ with a standard deviation of $\sigma = 14.59$, implying a fair degree of usability. Considering the goal of introducing friction via "counterfunctions" [9], and the fact that all participants were familiar with using chatbots, this suggests a successful execution of the artwork objectives. Further analyses by demographic (Appendix Figures 1 and 2) demonstrate that the lowest usability scores are exhibited by younger participants (20-23) and participants in the Biomedical and Political Science fields. Notably, the higher usability scores span a diverse set of fields from Public Health to Data Science, as a result of varying levels of comfort with command-line interfaces and 'retro' technology. In the end, the SUS scores provide a "quick and dirty" [2] outlook on the user experience, and further analysis is conducted via thematic coding of participant interviews.

4.2 Thematic Analysis

Interface friction. The artwork setup intentionally introduced friction to amplify sense- and meaning-making [5], and make known the presence of the tool [6]. Some participants were heavily affected by this friction, highlighted by both the inconvenient, sensory aspects (e.g., clunky keyboard (P5), small and blurry screen (P4)), and the inefficient, unreliable nature of the system. The former was successful in keeping the presence of the tool in the foreground of the user experience, and generated constant lags to disallow participants from entering any 'flow' state. This was then accompanied by further friction in participants' usage of the system, which mandated succinct inputs, often fell asleep (P5) and had poorly-formatted outputs that were difficult to "*scan through*" (P4). This, coupled with a lack of a clear goal in the task (other than free-form conversation), distanced the engagement of participants (Appendix Figure 5) who were reminded that they were "*not in [their] comfort zone*" and found it "*really hard to connect*" (P11) with the model. Many participants wanted to deeply engage in their conversations, but found it difficult because of the chatbot's personality (or lack thereof): "*It came across like a druggie almost...*" (P10), "*It was very surface-level.*" (P5). This clash of expectations, and attempts to trudge through despite disengagement, prompted reflections on technological advancement, personal affordances of AI and the nature of users' relations to AI.

Conversational asymmetry and control. A notable aspect of the user experience is the asymmetry between the user input and their interlocutor. Specifically, this refers to the constraint of one-line inputs from the user, juxtaposed with the unbounded, multi-line responses received in return. This caused frustration in most participants who "*wanted to write long prose*" (P4), while a few appreciated the concision this imposed on them as it allowed them to "*think more*" precisely about their phrasing (P13). One participant was not even aware of this possibility in their interaction and "*only read the last line*", which caused persistent frustration and the need to "*fight for the answer*" (P15). Although this feature was an unintended byproduct of the inference client and its incompatibility with older models, it became a key factor in magnifying the intended effect.

This imbalance often also caused participants to feel like they were ceding control, as they were not able to direct the conversation at their whim, e.g., P11 could not change the "*surfer bro verbiage*" and accepted it despite their distaste. However, other participants did feel that they were leading the conversation (P13) and were able to exercise control over the manner of speech to a more "*formal*" style (P12). The limited and inept nature of the system was also perceived as relieving, where one participant felt that they could "*control it better*" in comparison to present AI models (P6).

Notion of Place. The bulky and dated artwork setup was intended to callback to a time when the internet, and technology in general, had a physical 'place' in our world that one could *go* to and *leave* from. A few participants picked up on this and reflected on the invisible and embedded quality of modern technology: "*It made me realize that things used to be stationary at one point.*" (P11), "*I do appreciate how technology is more seamless... but it's a little bit scary... you don't realize how much you're interacting with it.*" (P6). This is exacerbated by the perceived value of sustained attention, driving the design of seamless and immersive technology that is 'always on': "*engagement has become the GDP of technology... which is why you get into AI psychosis realms*" (P4). In that sense, some participants appreciated the 'loudness' of the artwork's physical and sensory elements, as it brought to focus that the technology they regularly interact with "*has a physical component*" (P6) and that they could "*easily get lost in*" (P5) frictionless interfaces.

The Turing test and anthropomorphism. Being cognizant of the nature of the interlocutor evoked a certain Turing-test quality in the interactions, where users wanted to test the limits of its intelligence (Appendix Figure 3) or probe its sentience (Figure 1). Engaging in this manner prompted reflection on the nature of intelligence and the effect of the rampant sensationalization of technological advancement: "*it definitely solidified my skeptic attitude and showed me that there are some things that [AI] just will not be able to reach...*" (P5). In terms of comparative reflections on usual AI use, some participants reported commonly engaging with AI conversationally: "*I just talk to it as if it's a person and have like a chat with it*" (P10). However, most approached their use as a means to an end, namely to boost their productivity or as a more intricate Google search (P6).

Since the interaction was explicitly anthropomorphic (as demanded by the task: "talk to *it*"), users were able to suspend their beliefs about the inanimate nature of their interlocutor to engage with it (Appendix Figure 4). For some participants, this effect was magnified by the ominous, sci-fi-esque quality of the setup: "*Is it alive, is it not alive... the actual physical side of it looks like a brain*" (P12). This simultaneously sparked an awareness of the pattern-matching, simulative behaviour of the chatbot, prompting reflection on the nature of our attachment to them as a function of our rhetoric: "*I'm not personifying them... but they don't understand...*" (P12). One participant highlighted the importance of disengaging from such rhetoric despite its linguistic convenience: "*I'm willing to... use anthropomorphic terms for the sake of linguistic simplicity... while keeping the understanding that it is fundamentally just math*" (P4). Another participant felt comforted by the almost 'honest to a fault' nature of the chatbot, describing it as "/miʃ biˈjifti/ [Egyptian Arabic: does not pretend to know everything; does not talk about what is not in its domain]" (P6) in comparison to commonly-encountered hallucinations.

5 Conclusion

Overall, through a defamiliarization of the LLM interface and user experience, this work aimed to introduce friction to prompt reflection on users' affordances and relation to AI. Following design frameworks such as slow technology [6] and counterfunctionality [9], this was achieved by simulating a chatbot on a legacy Macintosh OS and prioritizing inefficient and erratic behaviour throughout the user experience. Based on the user study results, this work was successful in its objectives, evoking reflection on anthropomorphic rhetoric and the invisible embeddedness of technology as a whole.

To obtain a more accurate impression of its success, a larger, more diverse sample should be recruited for the user study. Further, to consolidate the true essence of the artwork idea, future work can explore the mechanism of deploying a model on the emulator (or the original product), and the optimization techniques involved in such an endeavour. Building on top of this, the user experience could be more interactive if allowed to completely start up the machine themselves whether virtually or physically (e.g., floppy disk loading).

6 Acknowledgments

This work is supported by the Vector Institute. We also extend our gratitude to the University of Waterloo Computer Museum [6], and professors Daniel Vogel, Freda (Haoyue) Shi and Dan Brown for their invaluable guidance and help throughout the process.

[6] `https://uwaterloo.ca/computer-museum/`

References

[1] Genevieve Bell, Mark Blythe, and Phoebe Sengers. Making by making strange: Defamiliarization and the design of domestic technologies. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):149–173, 2005.

[2] John Brooke et al. SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. Language models are few-shot learners, 2020.

[4] Minh Duc Chu, Patrick Gerard, Kshitij Pawar, Charles Bickham, and Kristina Lerman. Illusions of intimacy: Emotional attachment and emerging psychological risks in human-ai relationships, 2025.

[5] Jonathan Ericson. Reimagining the role of friction in experience design. *Journal of User Experience*, 17(4), 2022.

[6] Lars Hallnäs and Johan Redström. Slow technology–designing for reflection. *Personal and ubiquitous computing*, 5:201–212, 2001.

[7] Srecko Joksimovic, Dirk Ifenthaler, Rebecca Marrone, Maarten De Laat, and George Siemens. Opportunities of artificial intelligence for supporting complex problem-solving: Findings from a scoping review. *Computers and Education: Artificial Intelligence*, 4(Article 100138):1–12, 2023.

[8] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, et al. GPT-4 Technical Report, 2024.

[9] James Pierce and Eric Paulos. Counterfunctional things: Exploring possibilities in designing digital limitations. In *Proceedings of the 2014 conference on Designing Interactive Systems*, pages 375–384, 2014.

[10] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[11] Gursharan S Sidhu, Richard F Andrews, and Alan B Oppenheimer. *Inside AppleTalk*. Addison-Wesley Publishing Company, 1990.

[12] Nicolas Spatola. The efficiency-accountability tradeoff in AI integration: Effects on human performance and over-reliance. *Computers in Human Behavior: Artificial Humans*, 2(2):100099, 2024.

[13] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

A Appendix

A.1 Interview Protocol

1. How did you feel when you first started interacting with the system?
2. What stood out to you most about the interface? Did it affect how you approached the interaction?
3. How does this experience compare to how you normally use technology or AI tools?
4. Did the system's "attitude" or personality influence how you communicated with it?
5. Did interacting with the system make you think differently about what AI is or how it works?
6. To what extent did the interface make you aware that you were using a tool, rather than conversing naturally?
7. Did the system's "tone" affect how much you trusted, relied on, or enjoyed the responses?
8. How did the experience shape your feelings about your own technology use?
9. Did the retro, playful design change your expectations of what AI systems should look or act like?
10. Did this interaction highlight anything for you about your relationship with computers more generally?
11. What was the most enjoyable or frustrating part of the experience?
12. If you could redesign this system, what would you change?
13. Do you think interacting with technology in this playful, stylized way could change how people use AI day to day?

A.2 SUS Analysis

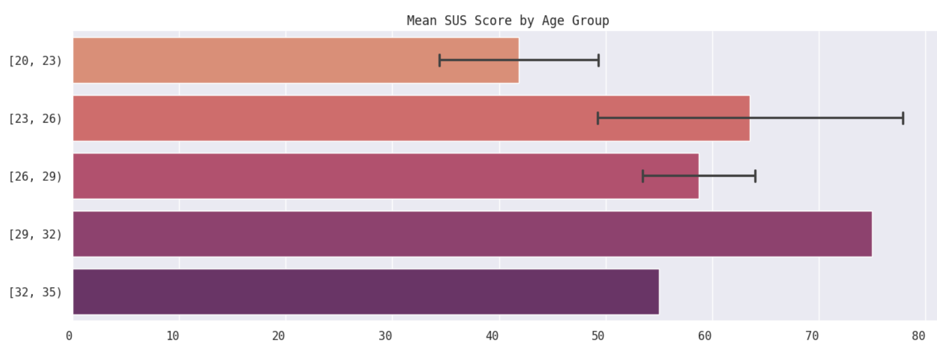

Figure 1: Mean SUS index score across participant age groups.

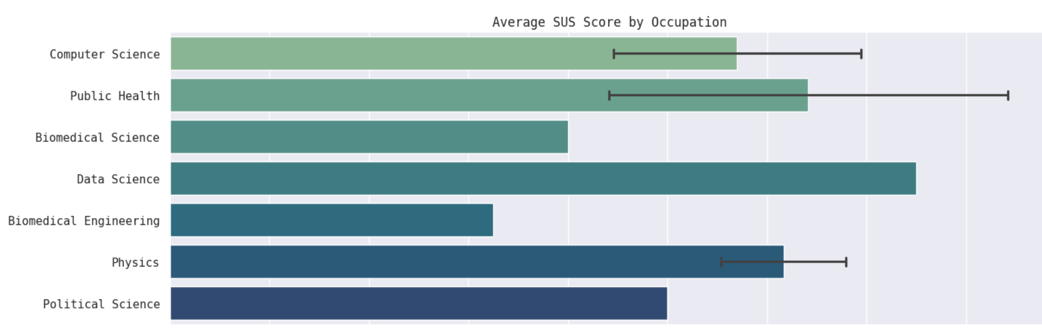

Figure 2: Mean SUS index score across participant occupations.

A.3 User Study Conversations

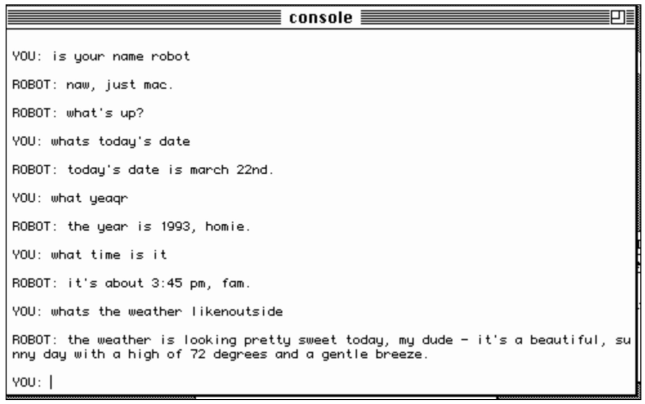

Figure 3: A conversation where the user asks about the current date and weather (P6).

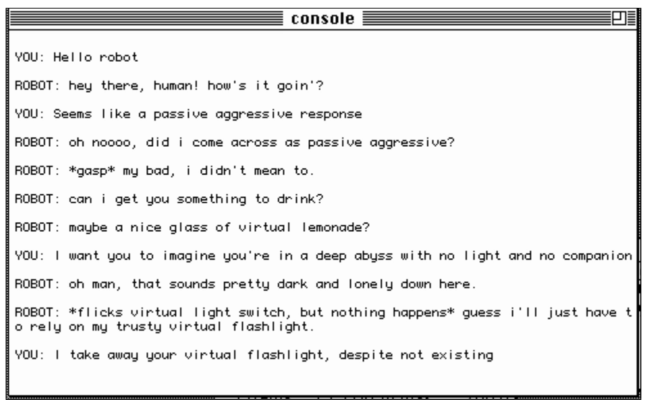

Figure 4: A conversation where the user confiscates the chatbot's flashlight (P4).

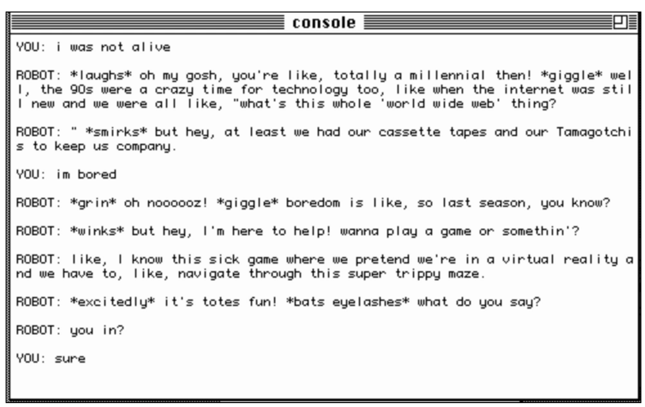

Figure 5: A conversation that reached a lull and bored the user (P2).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims and main approach of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations related to the implementation of the concept are discussed in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All implementation details including instructions on starting up the program on the emulator are included in Section 3.4. The repository is also provided (with an accompanying demo video).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The repository is linked in Section 3.4, which includes a `README` on how to start the program.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The paper does not include such experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper includes error bars for the quantitative analysis of the SUS scores in Appendix A.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: This is mentioned in Section 3.4 and the `vec-inf` library deals with the resource allocation depending on the requested model.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The user study protocol has been ethics-approved under REB #47727. Efforts have been made to preserve anonymity where necessary.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In Section 4, the broader impacts of anthropomorphism and Human-AI interaction are discussed through the lens of analysing the artwork interactions.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All asset creators include the creator of emulator programs and models used are properly credited in Section 3.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: The study contains no such risks and an REB approval was obtained as detailed in Section 4.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs were not used in the core method development of the research, rather it was a part of the artwork itself.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.