Poor Romeo: Final Report

Github

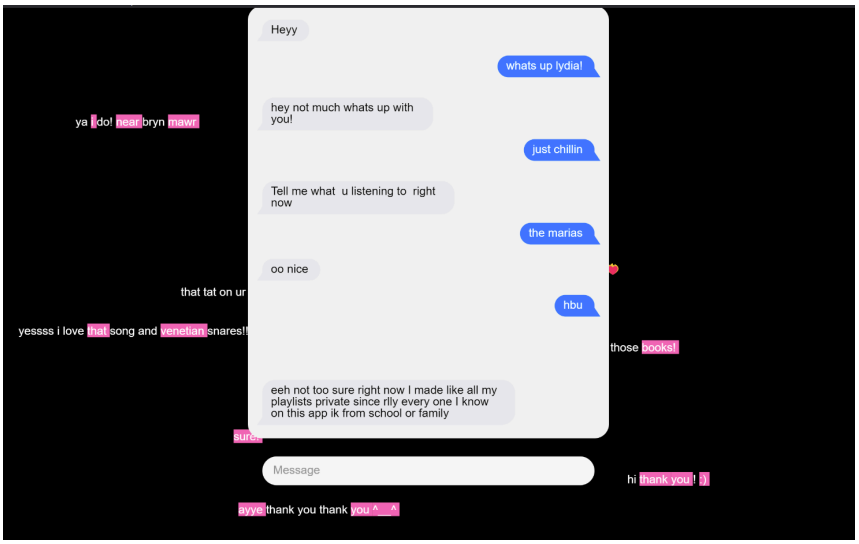
Noah Kornberg + Lydia Graveline
CART 498

Abstract

Poor Romeo is a real-time conversational experience that explores how social media and dating platforms collect and commodify personal data to construct digital versions of who we are, often without our consent. Using machine learning, we fine-tuned a language model on our own extracted data to generate chatbots that mimic our speech patterns, personalities, and conversational styles. The result is a speculative prototype of what these platforms already possess: a digital replica built entirely from our online presence. By inviting users to engage with these artificial selves, *Poor Romeo* questions the boundaries of identity, consent, and authenticity in the digital age. The project challenges the illusion that we control our online selves and reveals how much of "us" already exists in systems we don't own.

Introduction

In an era where personal data is continuously extracted, analyzed, and repurposed by digital platforms, the line between who we are and how we are represented online is becoming increasingly difficult to trace. *Poor Romeo* responds to this condition by asking a simple yet unsettling question: can a machine-generated chatbot, trained solely on the data collected from our social and dating app histories, convincingly mimic us? And if so, what does that reveal about the version of ourselves that already exists in the hands of platforms we don't control?



This project sits at the intersection of digital art, machine learning, and critical data studies. It engages with ongoing debates about surveillance, digital identity, and the politics of consent in online spaces. It builds on the speculative tradition of artists and theorists who examine how data becomes a proxy for selfhood, often in ways that bypass awareness, let alone permission.

Several projects inform our approach. Max Cooper's *On Being* uses real-time audience input to generate a dynamic portrait of human emotion, emphasizing the affective layers of digital interaction. Judith Donath's *Data Portraits* foreground the challenges of designing personal representation in virtual space, emphasizing the gap between who we are and how we appear when filtered through code and interface. *Life Sharing* by Eva and Franco Mattes—an early net-art example of radical self-surveillance—anticipated the total collapse of public and private that now defines online activity. Together, these works offer a foundation for *Poor Romeo*, which we could then further.

Our project also engages with recent academic discourse. Yoon (2023) warns of the ethical and psychological risks of AI clones, noting how machine-generated replicas based on scraped data blur the boundary between simulation and theft. Shoemaker (2023) explores how AI reshapes digital identity and its

security, arguing that biometric and behavioral data are becoming as central to personal representation as names or photos. These concerns underpin the stakes of *Poor Romeo*, which uses art to ask what it means to be digitally cloned before we've given permission to be replicated.

The significance of *Poor Romeo* lies in its ability to expose the extent to which our digital identities can be reconstructed without us. It does not offer a vision of the future—it reflects a condition already present. By generating a chatbot version of ourselves trained only on the data these platforms routinely collect, we invite participants to confront the version of "you" that already exists in algorithms, servers, and behavioral prediction models. It also complicates the idea of authenticity: if a machine can imitate your voice, mannerisms, and patterns of thought convincingly enough, what part of that interaction still belongs to you?

The project aims to achieve four main objectives:

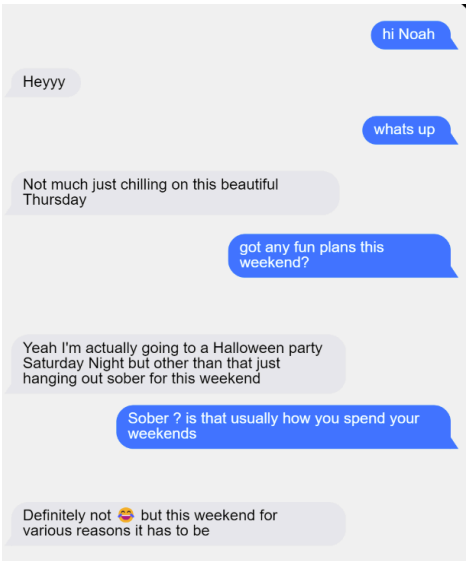
1. To demonstrate how little information is needed to create a digital proxy of a person.
2. To expose the ways in which digital platforms collect, replicate, and commercialize identity.
3. To question the meaning of consent, authenticity, and selfhood in platform-based interactions.
4. To provoke reflection on the future of personhood in an era of digitalization.

By merging machine learning with personal data and user interaction, *Poor Romeo* opens up a critical conversation about agency, representation, and what it means to be "you" in digital space. The creative vision behind *Poor Romeo* is to use generative AI as a speculative lens to explore how digital platforms construct and reproduce our identities without consent. This project does not treat AI simply as a creative tool. Instead, it questions how much of our identity already exists in the data we leave behind. By building chatbot versions of ourselves trained only on messages from Instagram and Hinge, we demonstrate how easily a digital replica can be created. These replicas simulate our speech patterns, preferences, and tone of voice. When users interact with them, the experience reveals how personal data can be repurposed into something both intimate and alien. The project challenges assumptions about authorship, privacy, and emotional authenticity in digital interactions.

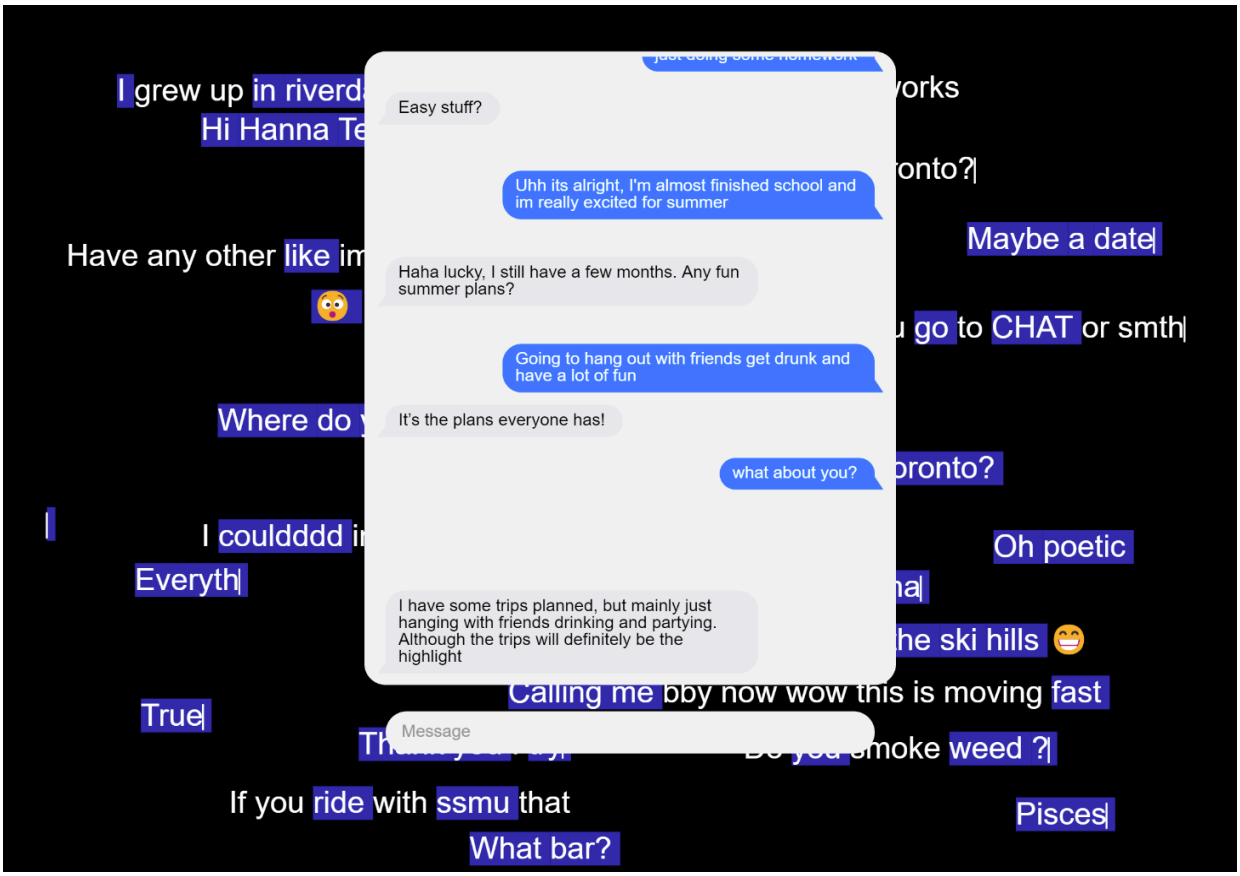
Technologies

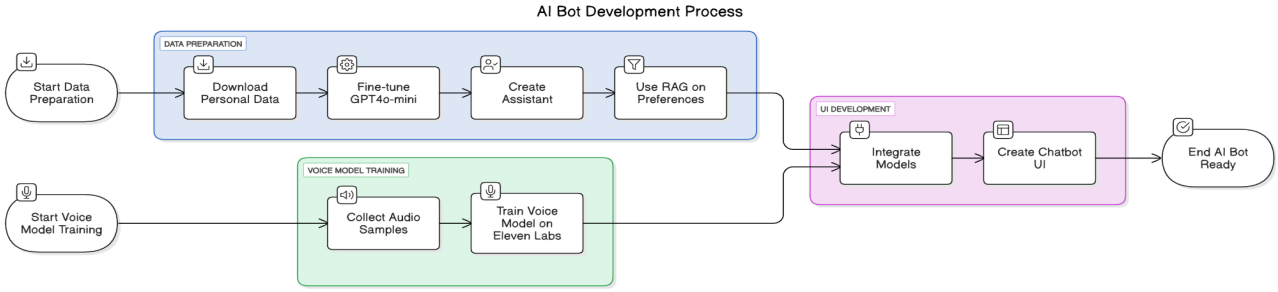
We used a combination of software tools and AI models to bring this idea to life. At the core of the system is OpenAI's GPT-4o-mini, which we chose for its conversational strength and cost-effectiveness. Using our personal message histories, we created two chatbots through fine-tuning GPT-4o-mini, prompt enginering, and RAG.

To provide additional context and specific facts about ourselves for our chatbots, we implemented retrieval-augmented generation (RAG). This method allowed us to provide personal preferences and frequently mentioned topics from our app profiles. For example, if someone asked the AI about pets or favorite music, the system could pull specific answers based on real data rather than generate generic replies. This approach made the conversations more grounded and personal.



The user interface was developed using p5.js. We chose to design it as a replica of iMessage to create a sense of familiarity and realism. The goal was to make users feel as if they were speaking with a real person through a platform they already use. We also added messages from our data in the background, so users could read the information the bot was trained on. This added another layer of complexity and vulnerability and we were displaying private messages for everyone to see. In addition to the text interface, we integrated voice synthesis to deepen the illusion of presence. We used ElevenLabs to generate cloned voices for each agent, based on five minutes of recorded speech. Each voice was created from short 30-second samples, which were used to produce natural-sounding responses. The AI reads its replies out loud using ElevenLabs instant voice cloning API, giving the impression of a live conversation with someone who sounds like us. Using ElevenLabs we created text-to-speech voice models trained on samples of our actually voices, matching the tone and rhythm of the original speaker.





Process

The development of *Poor Romeo* followed a structured yet iterative process. The project began with a clear vision for the user interface, inspired by Max Cooper's *On Being*, which emphasizes real-time, emotionally-driven user interaction. We began designing supplementary visual showcasing the data used to trained our models. At the same time, we submitted our data requests from Instagram and Hinge to retrieve our personal message histories; After requesting your data from a social media platform, it usually takes around 2-5 days for them to compile and send you the relevant information so while we waited for the data, we focused on interface development, laying the groundwork for how users would interact with the AI.

Once our data became available, we compiled and labeled it using a custom Python script. For creating our fine-tuned models, we used every message we've ever sent on our personal Instagram and Hinge accounts. These exports included a mix of messages—some useful, others not. To prepare the data for training, we developed a custom Python script that cleaned and reformatted the content. To emulate realistic, persona-based dialogue for OpenAI's fine-tuning format, we needed to clearly differentiate between messages written by us from those written by other users in the original conversation datasets. Each message was labeled with a "role" and "content" field to distinguish the speaker. Specifically:

- Messages that we wrote (i.e., our sides of the conversation) were labeled as "role": "assistant".
- Messages that others wrote (i.e., the people we were responding to) were labeled as "role": "user".

To do this accurately, we:

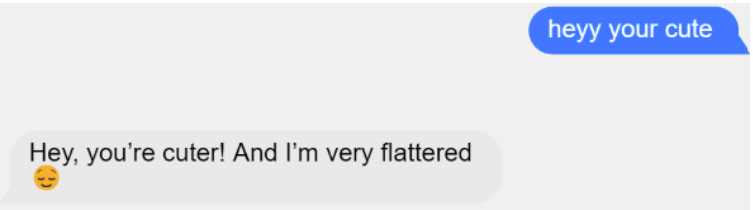
- Extracted sender names from each message.
- Defined our own name (e.g., "Lydia Graveline") as the assistant.
- For each message we sent, we created a new line of training data with "role": "assistant" and the message as "content"
- When applicable, we paired each of those assistant messages with the most recent message from the other user (i.e., the message sent immediately before ours), labeling it as "role": "user".

This pairing process allowed us to simulate natural, back-and-forth dialogue from real conversations. In cases where a preceding message from the other user was not available (e.g., at the start of a thread or in Hinge data), we used our message as a standalone assistant response. Unlike Instagram, the Hinge exports only include messages we sent and do not contain messages from other users. As a result, the Hinge data was used only for single-turn assistant messages, where no "user" message is present. These were still valuable for training because they provided stylistic, tonal, and content-related examples of how we communicate in romantic or flirtatious contexts.

This structured data was exported in the OpenAI fine-tuning-compatible JSONL format, with each line representing a self-contained conversation or message pair. In the end, Lydia's dataset included approximately 9,000 lines of conversation, mostly from Instagram. Noah's dataset contained around 5,000 lines, with more data coming from Hinge.

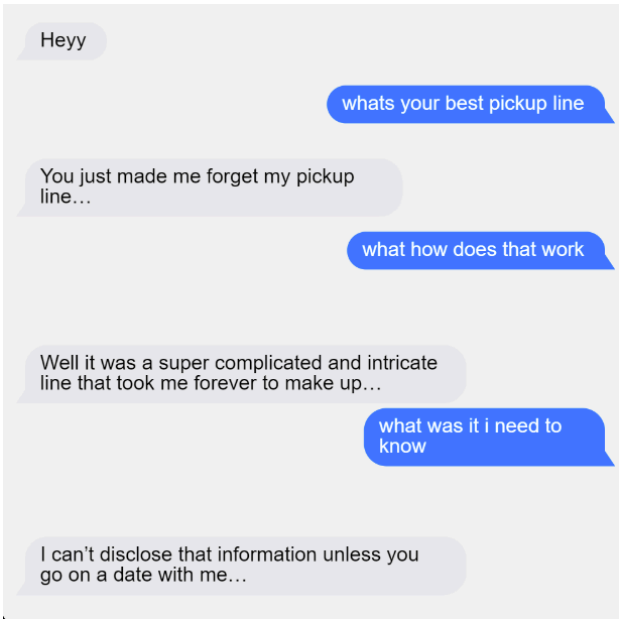
Our first attempt at fine-tuning using this data resulting with some unexpected issues: the model responded with strange symbols, failed to retain context, and often misinterpreted input. It seemed to be caused by issues with decoding special characters and emojis, causing weird glitches in the responses. Our messages, like most real digital conversations, used emojis frequently to express tone and emotion. However, the formatting and encoding of emojis became unstable as the data passed through different stages of the pipeline, from initial export to Python processing, training, and final output. Emojis were often reduced to odd character combinations like "DDs" or replaced with symbols that made no sense in context. This affected the realism and emotional nuance of the AI's responses. To fix this, we modified our script to remove non-conversational elements such as system notifications (e.g., "You started an audio call," "Reacted ❤️ to your message," etc.), emoji reactions, standalone emojis, and URLs using regex-based rules. By removing these lines, the user and assistant content pairings began to make more sense. Fine-tuning on this cleaner data produced much better results. We also found a way to reintroduce properly encoded emojis into the training set. This allowed the AI to send and respond to emojis in contextually appropriate moments, which significantly improved the personality of the interactions.

We experimented with the system prompt—the initial instruction that frames the AI's behavior. We gave it specific roles such as "be flirtatious," "act like you are on a dating app," or "be curious and seductive." These cues helped push the AI into more proactive conversational patterns. Unlike traditional AI models that tend to wait for user input, we wanted *Poor Romeo* to lead, to ask questions, and to simulate the charisma we often try to project in our real online interactions.



Another recurring issue was the inconsistency of the AI's memory. Despite feeding it accurate data, the AI sometimes confused key personal facts and hallucinated. At times, it landed close to the mark, producing responses that felt uncannily accurate. Other times it faltered, defaulting to repetitive replies like "hi", "uhhhh", "hahah", "idk" or losing track of the conversation entirely. During this phase, we discovered OpenAI's assistant framework, which allowed us to create persistent "threads" where the AI could maintain memory across an entire conversation. This resolved the memory issue and allowed us to foster more natural feeling conversations. This was a turning point in the project. Threads enabled a more natural flow, giving the chatbot the ability to reference earlier parts of the conversation and hold longer-term context—something standard GPT models often struggle with. Assistants also allowed us to easily implement retrieval-augmented generation (RAG) using vector spaces to provide personal preference data collected from our dating profiles, making the agents more consistent and relatable. However, despite using RAG, our bots still have difficulty retrieving specific facts like zodiac sign or height accurately. For example, it once identified a Virgo as a Capricorn, even though the correct

zodiac sign was clearly stated in the vector data. This revealed the limits of how much the model could internalize from a relatively small dataset. In practice, the AI existed on a sliding scale between ChatGPT's general tone and a fully personalized version of ourselves.



Despite the technical obstacles, the project yielded several successes. The integration of real voice samples using ElevenLabs gave the chatbot a stronger sense of presence. The iMessage-style interface made interactions feel familiar and emotionally grounded. Some conversations unfolded in surprising ways, capturing both the tone and rhythm of our personalities with unsettling accuracy. These moments highlighted the potential of generative AI not just to imitate, but to embody, digital versions of ourselves. Throughout this process, we learned to balance technical constraints with creative intent. We gained practical experience in data formatting, model fine-tuning, vector embedding, voice synthesis, and UI design. We also confronted the limitations of generative systems when tasked with replicating something as fluid and contextual as a human personality. More importantly, we discovered how these systems can be steered, shaped, and framed through careful prompting, structured input, and thoughtful interface design.

Future Work

While *Poor Romeo* successfully demonstrates the potential of generative AI to simulate digital versions of ourselves based on limited personal data, there are several areas for expansion and improvement. Future iterations of the project would aim to refine the fidelity of each digital replica by increasing both the quality and variety of input data. In this version, we worked primarily with message histories from Instagram and Hinge. However, a more accurate simulation could be achieved by incorporating data from a broader range of platforms. Adding message histories from other apps, such as WhatsApp, Messenger, Snapchat, or even email, would provide a more comprehensive understanding of our language, tone, and behavioral patterns across different contexts. This would allow the AI agents to generate responses that better reflect how we adapt our communication based on the platform, audience, or topic.

Alongside expanding the dataset, we would also explore re-tuning the model with a larger volume of labeled inputs and additional conversational cues. This would help improve the accuracy and fluidity of the chatbot's responses. Increasing the number and diversity of voice samples would also enhance the voice synthesis component, allowing the AI-generated voice to more closely capture the subtleties of our real voices. This includes emotional intonation, pace, and inflection, contributing significantly to the perception of authenticity

in spoken interaction. Re-examing our RAG data and the bot's ability to retrieve specific facts and details about our persona is another area for improvement.

Beyond technical refinement, we are also interested in testing the digital agents in live, real-world environments. One potential direction is to deploy the AI versions of ourselves into actual dating platforms to observe how they would engage with others. This raises provocative questions about consent, deception, and emotional labor. Would these bots be able to maintain meaningful interactions? How would others interpret or respond to an AI-generated personality? This experiment would open up a critical space to examine the ethics of automation in intimate settings, especially when the line between human and machine is no longer obvious.

By continuing to scale and refine the system, *Poor Romeo* could move closer to producing a digital self that is not only convincing but also interactive in complex social environments. Future iterations would not only sharpen the technological implementation but also deepen the project's critical engagement with the broader cultural and ethical implications of identity replication in the age of generative AI.

References

Cooper, Max, Ksawery Komputery, and Minjeong An. On Being. Accessed April 10, 2025. https://www.maxcooper.net/on-being.

Donath, Judith. 2014. The Social Machine: Designs for Living Online. Cambridge, MA: MIT Press.

Eva and Franco Mattes. Life Sharing. 2000–2003. Net Art Anthology. Accessed April 10, 2025. https://anthology.rhizome.org/life-sharing.

Shoemaker, Phillip. 2023. "The Role of AI in Enhancing Digital Identity Security." Identity.com, August 10, 2023. http://identity.com/the-role-of-ai-in-enhancing-digital-identity-security/.

Yoon, Donwook. 2023. "AI Clones Made from User Data Pose Uncanny Risks." The Conversation, June 4, 2023. https://theconversation.com/ai-clones-made-from-user-data-pose-uncanny-risks-206357.