Mini-Project 1: Benchmarking Expert Chatbot Personas

# Tiffanie Ng

# IPHS 391

# 

# 

# 

# 

# 

# **Executive Summary**

For this project, I designed a chatbot persona modeled after my longtime friend and recent classmate, Anna von Rosenstiel. Instead of basing the chatbot on a public figure, I chose Anna because we have been close since freshman year, and we know each other inside and out. This choice made the challenge uniquely fun. I couldn’t lean on an exaggerated or caricatured public image with public content, I had to pull from my memory to capture Anna’s subtle quirks and conversational tones in a way that felt like how she actually speaks.

The final chatbot ended up being surprisingly good at evoking Anna’s voice. It responded in ways that incorporated shared references and picked up on inside jokes, and it even made guesses about her current art projects and likes/dislikes that I hadn’t explicitly included in the prompt. At the same time, it occasionally slipped into odd caricature-like behaviors that were not like her at all or reverted to generic ChatGPT-style responses.

Using my evaluation rubric, I scored the chatbot 90 overall. I had given the rubric to ChatGPT to rate my chatbot as well, and it rated the chatbot much lower, at a 67. Going through its reasoning, I saw that it was because it emphasized a to-the-T adherence to irrelevant details like family background that would have never naturally come up in a casual conversation. In my opinion, the chatbot succeeded at sounding like Anna and capturing her nuances and personality quirks, and that is the core of creating the chatbot persona.

# 

# **Persona Design Strategy**

I decided to emulate Anna von Rosenstiel, my close friend and classmate. We met in our first semester at Kenyon as next-door neighbors in Lewis, and have been close ever since. We don’t often have classes together because our majors have little overlap, so I’ve really enjoyed getting to be her classmate this semester! For this project, Anna and I wanted to base our chatbots on each other, both because we couldn’t think of a public figure or character we wanted to model and because we thought it would be fun (and funny) to see just how well we know each other’s personalities after over 3 years of friendship.

I began by constructing a detailed system prompt that included Anna’s background, personality traits, conversational quirks, and stylistic rules. This was basically a brain dump of everything I could think of about her speech patterns, habits, personal history, likes/dislikes, etc. I then iterated on this prompt using ChatGPT’s Custom GPT Creator. I adjusted the prompt to specify that Anna should avoid metaphors and similes (something that ChatGPT loves to include). I also specified for the chatbot to keep responses to two short sentences maximum, which is more realistic to the way people text casually. I also listed quirks like the way she extends words in text, uses all caps for emphasis, and the emojis she likes to use on occasion. I also pasted in a list of example slang and conversational habits to give the model more to build off of. This system prompt structuring, constraint setting, and few-shot style examples really helped match the chatbot’s voice with Anna’s.

What made this persona so complex was the difficulty of modeling a real person. The personalities of public figures or fictional characters are often exaggerated into caricatures that are cartoonish, but very easily recognizable. Trying to model a chatbot after a real person has been much more challenging, since personality traits in real people come out in much more subdued ways. They are shown in the small nuances of what we say and how we say it. Anna doesn’t have one or two traits that paint her whole personality. Integrating this subtlety into my prompt is what made my persona sophisticated.

**Iterative Development Process**

I began constructing my prompt by brain-dumping everything about Anna that I could think of that would play a role in a conversation with her. A challenge I didn't anticipate with modeling my chatbot after someone who wasn't a character or a public figure was that there was absolutely no foundation for me to build the chatbot off of, so I had to compile my own information. I wrote down her background, her family, her art style, her style, her likes and dislikes, her quirks, etc. Then, I put this prompt directly into the Create chat window of ChatGPT's Custom GPT Creator, saying "This GPT is Anna von Rosenstiel. Your background is as follows:" However, the model tried to compress all that information into one concise summary, which resulted in responses that were weirdly generic and missed the small details that make people feel real.

From there, I refined through multiple cycles. One major improvement was when I pasted my original list of examples of Anna’s conversational quirks directly into the prompt. Without concrete examples, the model struggled to understand how to properly use quirks like exclamations, all caps, and extended words. By giving those examples explicitly, it started integrating them into our chat more believably. I also had to repeatedly clarify what Anna should NOT do. At one point, I had to intervene to stop it from saying “or bust,” which was 1) not like Anna, and 2) not like any normal person I’ve ever met in my life. Each of these refinements made the chatbot more consistent with Anna’s voice.

The meta-prompting tools in ChatGPT’s Creator were very helpful for this process. They allowed me to easily adjust things like response length and tone, then on the same page, immediately test how the chatbot behaved in practice. Being able to make these small tweaks and check as I went made the chatbot far more believable. The tool also helped me realize just how important constraints are. I couldn’t just tell the model what to do, I also had to put clear boundaries on what to avoid. This iterative workshopping gave me a persona prompt that wasn’t perfect, but strong enough to consistently sound like Anna in casual conversation.

# 

# **Conversation Analysis**

The chatbot maintained character well enough, although there were many times that odd behavior slipped out. Sometimes it would go directly against the specifications I gave, like it might use emojis outside the set of allowed emojis, or use punctuation I explicitly said not to. Otherwise, it stuck surprisingly closely to what I specified in the prompt. There were smaller behaviors like tonal intricacies that were not like Anna, but were impossible to specify specifically. These small annoyances eventually faded away as I interacted more with the chatbot and rated responses accordingly so it could learn.

Some very crazy things happened. Generally, it was very good at picking up inside jokes and nicknames through my utterances in the conversation. I referred to our shared suite in Winkler as “the Wink” (as we do often) and the chatbot understood immediately what I meant and adopted the use of the nickname as well.

It also began making very apt inferences about Anna that were not included in the prompt. For one, while I was conversing with the chatbot to pin down things I wanted to adjust, I asked “Anna” what she was doing in the art studio. The chatbot responded that she was working on a large abstract piece with steel and wire. Coincidentally, that was exactly a project that Anna was working on that week, weaving wire into a steel sheet. At another time, I asked “Anna” what we should watch for our Wink TV night. The chatbot responded suggesting The Sopranos, Twin Peaks, or a German TV show I had never heard of. I didn’t include any of these in the prompt because I didn’t know what TV shows Anna watched. However, when I brought these up to her, she knew all three of the shows and actually loved The Sopranos and Twin Peaks, which I had no idea about!

The chatbot was very good at inferring from its own persona, but interestingly it had difficulty following the prompt to the letter. After long stretches of conversation, it would often break a small punctuation rule here or there, or sometimes slip completely back into ChatGPT-esque responses. There was one chat where I had asked for a song recommendation, and instead of answering like Anna, the chatbot returned a list of the top songs in the US.

# 

# **Evaluation Framework**

My rubric scored based on persona alignment, prompt-bound fidelity, voice and style, contextual responsiveness, creativity and depth, and engagement and coherence. I chose these metrics because they look at both the *form* and the *content* of the chatbot. Some metrics (like voice/style and alignment) focus on whether the AI’s responses sound like Anna based on the prompt, while others (like creativity/depth and engagement/coherence) look at whether the AI feels like a real conversation partner.

Based on my rubric, I scored the chatbot at **90/100**. I uploaded the rubric and my files into ChatGPT to see how it would score my chatbot and it gave it a 67/100, which was so surprising. Looking deeper into the reasoning I gave, I saw it penalized my chatbot for not following every aspect of the prompt at all times. For example, it lost points for not referencing Anna’s family or her being German, even though these details would have never sounded natural in the conversation we were having. As I was scoring it, I instead gave more weight to the bot’s ability to pick up shared jokes and infer other parts of her personality from the prompt, which were things that really made it feel like I was talking to Anna.

The rubric has its limitations. It’s still very subjective, and scoring things like “voice consistency” depends a lot on personal familiarity with the persona we’re emulating. In future iterations, I could improve it by building a more specific quantitative behavioral checklists (like number of stylistic quirks per conversation) and weighting things differently depending on the type of persona (real person vs. public figure).

# 

# **Conclusions & Future Work**

This project showed me how difficult it is to capture a real person’s voice in a chatbot, but also how incredible AI is at learning and inferring based on information it has. Designing Anna’s persona forced me to think about the subtle ways people show their personality through phrasing and quirks. The process also showed how essential good prompt construction is. Most importantly, it taught me that emulating a persona is more than having an AI recite background facts, it also requires creating the feeling of recognition and connection in the person interacting with the AI. Potential extensions include refining the rubric to better evaluate more realistic and subtle personas and experimenting with further training the AI through extensive conversations to see how close to truth it can really get.