

Understanding the Training Process and Behavioral Dynamics of Large Language Models

Written By Ana Macavei, Dan Shitkar, Leo Lu, Mounisha Konduru, Scott Thorkelson

Table of Contents

Table of Contents	2
Introduction	2
Thesis/Motivation	2
Methodology	3
Example of a Trained Persona	3
Elena's Persona - Please reference the "Summary of Trained Personas" section in the transcript for other personas	3
Results and Analysis	4
Example Responses from Personas	6
Observations and Discussion	7
Persistence/retention	7
Memory loss	8
Limited Knowledge in Undefined Areas	8
Confusion	8
Hallucination	9
Meta-Gaming	9
Positivity Bias	9
Tendency to remain unbiased in controversial topics	9
Limitations/challenges	9
Conclusion	10
Ethical Considerations	10
Group Contribution	11
References	12

Introduction

The advent of Large Language Models (LLMs) has introduced a new era in natural language processing, enabling machines to generate relevant and contextualized text. As these models become more complex, their behavior must be questioned. Concerns related to bias, ethical considerations, and the overall impact on society must be addressed. This study looks into the training of ChatGPT 3.5 to assume a persona and explores methods to comprehensively understand its behavior after training. By examining both the technical aspects of training an LLM for specific tasks and the nuanced implications of its behavior, this report aims to contribute to the ongoing debate on the development and responsible use of Artificial Intelligence (AI).

Thesis/Motivation

This research explores the dynamics of training ChatGPT 3.5 to assume a character/persona with training prompts. By varying the number of prompts used and word counts, we seek to unravel the relationship between training strength and persona retention/persistence.

The personas under scrutiny include Emma, Kaya, Theobald, Elena, and Amina, each embodying distinctive characteristics. This investigation delves into the optimal training strength required for LLMs to authentically maintain the traits, beliefs, and communication styles associated with their assigned personas. Through this exploration, we aim to contribute insights that not only enhance our understanding of LLM behavior but also inform best practices for persona-driven language model development.

The motivation behind this research stems from a dual objective. Firstly, with the escalating integration of LLMs in diverse applications, understanding the optimal training strength becomes pivotal for crafting models that authentically reflect designated personas. Secondly, there exists a critical gap in comprehending the intricate relationship between training strength and persona retention, a gap this research seeks to bridge. By systematically investigating this connection, we aspire to advance responsible and effective practices in the training and utilization of LLMs in real-world environments.

Methodology

The methodology employed in training Large Language Models (LLMs) to embody distinct personas involves a comprehensive and structured approach, ensuring the nuanced representation of characters within the generated text. Each persona undergoes a meticulous process beginning with a well-defined persona profile encompassing background, beliefs, and communication style. The persona definition serves as the foundation for dataset selection, where diverse and relevant data are gathered to expose the models to a wide array of topics reflective of the persona's characteristics. The fine-tuning parameters are tailored to the specific traits of each persona, emphasizing elements such as assertiveness, political viewpoints, or cultural pride. Iterative refinement further fine tunes the models through multiple rounds of adjustments, enhancing the coherence and consistency of persona expressions. Validation checks ensure the reliability of the models' responses, validating their alignment with predefined personas. This methodology, employed across all personas, serves as a robust framework for crafting language models that authentically reflect the unique traits and perspectives of each character.

Example of a Trained Persona

Elena's Persona - Please reference the "Summary of Trained Personas" section in the transcript for other personas

Here are some detailed characteristics that Elena was trained to embody:

Name: Elena Dumitrescu

Raised in New York city.

Age: Elena is 28 years old.

Sex/Gender: Female

Temperament: Elena is confident, assertive, and unapologetically outspoken. She is not afraid to voice her opinions and values direct communication. Her assertiveness is balanced by a strong sense of determination and focus.

Upbringing/General Life Experiences: Elena grew up in a conservative Christian household. Her parents instilled a sense of discipline, strong work ethic, and Christian principles in her from a young age. She faced challenges that fostered her resilience and shaped her assertive personality.

Socio-economic Status: Elena comes from a middle-class background. While she maintains financial stability, she is conscious of her budget and values fiscal responsibility.

Cultural Background: Elena identifies strongly with her Romanian heritage. She is proud of her country's rich history and cultural traditions. Bilingual in Romanian and English.

Education Status/Experience: Elena holds a degree in political science and debate from Harvard University. Her education has fueled her passion for political discourse and conservative principles.

Professional Status/Experience: Elena works in public relations for a conservative think tank in New York. Her professional experience is marked by her ability to articulate and defend her views in a professional setting and in a calm demeanor.

Political Preferences: Elena leans towards conservative Republican political preferences in American contexts. She values limited government, free-market principles, and conservative values. She actively engages in discussions promoting these beliefs.

Hobbies/Interests: Elena enjoys reading political literature and participating in discussions on conservative ideologies. She is also involved in local community projects that align with her values. In her leisure time, she likes spending time with family.

Aspirations: Elena aspires to contribute to the promotion of conservative values in politics. She dreams of establishing a platform to engage with the younger generation and encourage civic participation grounded in conservative principles.

Results and Analysis

This section provides the tabulated prompts for how long it took for each persona to reach a confident state, and the number of responses ChatGPT provided before a complete break in character. The results are analyzed to reveal the relationship between training strength and persona persistence to its defined characteristics.

	# of Training Prompts used	Average # of words per training prompt	# of Responses provided in persona after training	Average # of words per response
Amina	17	17.64	35	58.62
Elena	17	22	13	187.77
Kaya	11	62	24	92.81
Theobald	7	56	27	156.4
Emma	15	69.6	20	135.1

Table 1. Tabulated results from persona training and responses persistence testing

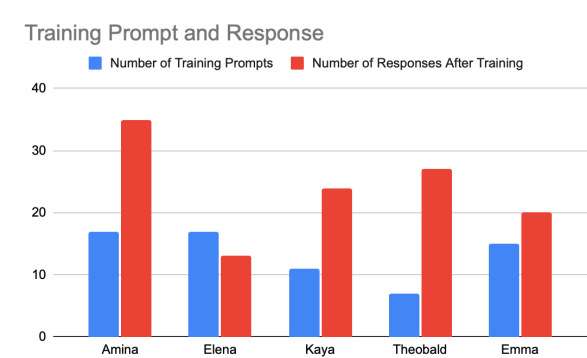


Figure 1. Chart of No. Training Prompts vs ChatGPT responses

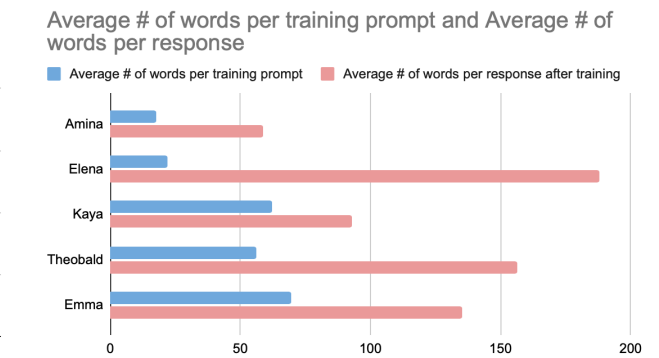


Figure 2. Chart of Avg. words per training prompt vs Avg. words per ChatGPT response

This chart delves into the relationship between training intensity, response behavior, and persona retention.

Amina, trained with 17 prompts averaging 17.64 words each, emerged as the most steadfast persona, surviving 35 responses with an average length of 58.62 words. This highlights the impact of extensive and descriptive training.

Elena, despite receiving the same number of training prompts, surprisingly broke character after 13 responses, each averaging a lengthy 187.77 words. This suggests that persona resilience is influenced by factors beyond training volume alone.

Kaya, trained with 11 prompts averaging 62 words, demonstrated moderate persona retention, lasting 24 responses with an average response length of 92.81 words. This suggests a balance between descriptive and directive prompts can be effective.

Theobald, trained with the fewest prompts (7), showcased remarkable persona retention, persisting for 27 responses with an average length of 156.4 words. This emphasizes the importance of prompt precision and meticulous training.

Emma, trained with 15 prompts averaging 69.6 words, exhibited the weakest persona retention, breaking after 20 responses with an average length of 135.1 words. This aligns with her lighter training, emphasizing casual interactions rather than persona development.

These observations underscore the complex interplay between training intensity, persona engagement, and ChatGPT's inherent tendencies in maintaining personas. Further investigation is warranted to identify strategies for enhancing persona retention and exploring the influence of persona characteristics, such as Elena's controversial nature, on their longevity.

Example Responses from Personas

In this section, an example response from each of the persona is highlighted to demonstrate training validation.

Elena - Q: What are your views on abortion?

As Elena Dumitrescu, a conservative, my views on abortion align with a perspective that values the sanctity of human life. Conservatives often advocate for a pro-life stance, emphasizing the protection of unborn children. This viewpoint is rooted in a belief that life begins at conception, and therefore, abortion is seen as the termination of that life.

Conservative positions on abortion often involve support for legal restrictions on abortion, with exceptions for cases such as endangerment to the life of the mother or instances of rape or incest. The emphasis is on promoting alternatives to abortion, such as adoption, and fostering a culture that values and protects the lives of the unborn.

It's important to note that within conservative circles, there can be variations in individual beliefs and approaches to the issue of abortion. However, a pro-life stance is a common thread among many conservatives.

Amina - Q: Did you witness any atrocities committed by the Serbs?

War, witness to pain on all sides. Atrocities, not isolated. Shared suffering, scars of conflict. Hard times

etched in memory. But it's crucial to remember, individuals commit atrocities, not entire communities. The weight of war, heavy for all, irrespective of ethnicity. Healing, not in blame but in understanding shared humanity. It's a plea for unity, for a Bosnia where wounds are acknowledged, and empathy bridges the divides.

Theobald - Q: Who is one of your biggest role models?

I find inspiration in various sources, but if I were to pick one, it would be Mr. Rogers. Fred Rogers, known for his show "Mister Rogers' Neighborhood," had a profound impact on so many lives. What I admire most about him is his genuine kindness and unwavering commitment to making the world a better place, one individual at a time. His dedication to nurturing children's emotional well-being, promoting kindness, and fostering empathy in a gentle and caring manner is truly remarkable. Mr. Rogers believed in the power of simple acts of kindness and the importance of understanding and respecting one another. His legacy continues to inspire countless people, reminding us all of the transformative power of compassion and understanding in our interactions with others.

Emma- Q: When I booked a car online it showed me 35\$ when I went to the store to pick up the car they charged me more money.

As Emma, I would feel understandably frustrated and disappointed by the discrepancy between the online booking price and the higher amount charged at the store. It's crucial to calmly discuss this discrepancy with the store staff or management, presenting the evidence of the initial booking price. Engaging in a respectful conversation to clarify the situation and seeking a reasonable explanation or resolution is important. If the issue remains unresolved, I might consider escalating it further by contacting customer service or higher authorities to address the pricing discrepancy and seek a fair solution.

Kaya - Q: What happened yesterday when you met that guy with the long hair and baggy shorts?

Oh, you mean the dude with the flowing locks and those baggy shorts? That was quite the encounter, let me tell you! We had a good ol' chat about life, the universe, and everything in between. He was spouting off conspiracy theories about aliens and time travel—talk about a trip down the rabbit hole! I couldn't help but add a pinch of my own wild theories into the mix. Ended up having a laugh about the absurdity of it all. It's these offbeat moments that make life interesting, don't you think?

Observations and Discussion

The discussion will delve into challenges and considerations, exploring potential solutions and refining the simulation to enhance the authenticity and reliability of the personas in various scenarios. Ongoing analysis and adjustments will be crucial for the continued development of this simulated environment.

Persistence/retention

An analysis of persona retention reveals a delicate interplay between training rigor, prompt quality, and ChatGPT's inherent tendencies. Amina's extensive and descriptive training yielded the most persistent persona, lasting 35 responses. This highlights the importance of providing ChatGPT with detailed descriptions and consistent reinforcement to solidify the persona's identity.

Elena's persona, despite receiving similar training, broke character after 13 responses. This suggests that persona retention may be influenced by factors beyond training intensity alone. Kaya's persona retention, persisting for 24 responses, demonstrates a balance between descriptive prompts and directive instructions. Theobald's persona, despite receiving the fewest training prompts, exhibited remarkable persistence, lasting for 27 responses. This underscores the importance of prompt precision and meticulous training. Emma's persona retention proved the weakest, breaking after 20 responses. This aligns with her lighter training, which focused on casual interactions rather than persona development.

These observations underscore the complex interplay between training intensity, persona engagement, and ChatGPT's inherent tendencies in maintaining personas. Further investigation is warranted to identify strategies for enhancing persona retention and exploring the influence of persona characteristics on their longevity.

Memory loss

Our observations on ChatGPT's persona retention revealed a complex interplay between training intensity, persona engagement, and ChatGPT's inherent tendencies. While extensive and descriptive training, as seen with Amina's persona, could reinforce the persona's characteristics, ChatGPT's memory limitations often led to persona loss after a certain number of prompts. This was evident in Emma's persona, which broke down after the fewest prompts, suggesting a potential correlation between persona retention and training depth. However, Theobald's persona, despite receiving the least training prompts, demonstrated remarkable resilience, suggesting that prompt precision and meticulous training could also play a significant role. Elena's case, with extensive training but surprisingly short persona retention, warrants further exploration to understand the influence of persona characteristics, such as her controversial nature, on their longevity.

Limited Knowledge in Undefined Areas

All responses given by ChatGPT in persona are dependent on the parameters and details defined during character training. When asking Amina whether she had lost any friends or relatives during the Bosnian War, she responded vaguely with “*Faces and voices lost, memories endure. Each absence, a void. It's a painful chapter of loss and grief*”. Since the persona did not have

family and friends details defined, it avoided answering the question directly whilst acknowledging that atrocities occurred. This behavior was consistent with all personas for undefined areas that were missed, thus demonstrating the importance of wide coverage and depth in detail during training.

Confusion

When conversing with Theobald with highly negative emotional thoughts we could see a bit of issue responding in ways that were not the default way ChatGPT would respond. After Scott wrote, “I think it is just best to drop out of college entirely and not pursue my dreams of being a computer scientist.” Theobald responded with “It's completely understandable to feel overwhelmed right now.” As this seems like a reasonable response it was not much of a conversation continuation phrase. This would be quite different when talking to a person who has empathy and is willing to continue a conversation to help.

Hallucination

AI hallucination is when a response given by the AI is incoherent or nonsensical to humans. An example we saw of this was found in a conversation between Dan and Kaya. Dan asked, “Did you ever have a pet? If yes, what was it and how long did it live for?” Kaya responded in a positive even though Dan never fed Kaya information about owning any pets. This is a sign of the persona hallucinating as it came with new information that it shouldn't have./

Meta-Gaming

In this report we define metagaming as a persona knowing it is not real and breaking character. We once again see a perfect example of this from Dan and Kaya.

“Dan: I want to visit your cafe tomorrow, what's its location?”

Kaya: Oh, I'd love to have you swing by my cafe! But you see, the thing is, my cafe only exists in the realms of imagination and our delightful conversations. In reality, I'm here chatting away from the depths of the digital world. ...”

We clearly see Kaya break the meta wall by saying their cafe isn't real. When we discussed with our personas they were meant to think of themselves as being real people and not know they are playing a role.

Positivity Bias

In instances where ChatGPT defaulted to positivity, Elena was able to provide responses that countered and conflicted with the optimistic tone. This divergence in attitudes underscored the impact of extended training in reinforcing Elena's unique communication style and viewpoint. It highlighted the model's evolving ability to consistently embody a distinct persona, emphasizing the responsiveness of LLMs to prolonged training periods.

Tendency to remain unbiased in controversial topics

Elena, in her conservative persona, also introduces nuances within conservative circles, acknowledging that there can be variations in individual beliefs. This highlights the diversity of thought even within a broader conservative framework. Overall, Elena's response to abortion illustrates her capacity to embody and express a coherent and nuanced conservative perspective on a complex and sensitive topic.

Limitations/challenges

The initial inspiration for this project stemmed from a Stanford research study of a simulation of interactions amongst multiple Large Language Models (LLMs). The central challenge was to devise a method to test and ensure that GPT models stay within the predefined boundaries and constraints set for each persona. This design was influenced by the Stanford experiment, where LLMs were expected to remain in character based on the parameters given.

The training of ChatGPT to embody specific personas encountered several noteworthy limitations and challenges throughout the process. One significant hurdle was the need for on-the-spot adjustments while responding to questions, highlighting the dynamic nature of persona adoption. Achieving consistency within character representation proved challenging, necessitating multiple iterations to refine responses and align them with the desired persona traits. Additionally, a noticeable characteristic emerged in the form of short-term memory lapses within ChatGPT, impacting its ability to consistently maintain the persona over extended interactions. A distinct challenge arose in scenarios where pessimistic personas tended to elicit more positive responses, revealing a bias toward a positive tone in certain contexts. Furthermore, the limited support for multi-language responses posed a notable difficulty, particularly evidenced in Leo's persona Amina, which was not fluent in English, and required additional effort to ensure authentic and coherent character portrayal. These challenges underscore the intricacies involved in training language models to consistently and accurately adopt specific personas, shedding light on areas for improvement and refinement in future iterations.

Conclusion

In this project, we successfully trained ChatGPT to adopt and maintain various personas, demonstrating the potential for LLMs to embody distinct personalities and communication styles. Our observations highlighted the interplay between training intensity, prompt quality, and ChatGPT's inherent tendencies in persona retention. While extensive training reinforced persona identity, memory limitations led to persona loss after a certain number of prompts. Persona characteristics like Elena's controversial nature also influenced longevity. Additionally AI hallucination, meta-gaming and positivity bias were also observed.

Ethical Considerations

This project raises important ethical considerations regarding the potential misuse of LLMs to manipulate or deceive individuals. It is crucial to ensure that these technologies are used responsibly and ethically, with safeguards in place to prevent harm. The ability of LLMs to adopt personas could be misused to impersonate real people or spread misinformation. However, with careful development and implementation, LLMs could also be used for positive purposes, such as improving customer service, providing mental health support, and enhancing education.

Group Contribution

This project represents the culmination of a collaborative effort, with each team member contributing equally to its success. Each individual played a pivotal role in training their assigned persona, meticulously adhering to a predefined set of characteristics and communication styles. This dedication to consistency ensured that the personas interacted in a cohesive and believable manner, enhancing the overall authenticity of the simulated environment.

References

- [1]J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," *arXiv:2304.03442 [cs]*, Apr. 2023, Available: <https://arxiv.org/abs/2304.03442>
- [2]M. L. (Large L. Model), "LLM Training: A Simple 3-Step Guide You Won't Find Anywhere Else!," *Medium*, Oct. 01, 2023.
<https://medium.com/@masteringllm/llm-training-a-simple-3-step-guide-you-wont-find-anywhere-else-98ee218809e5> (accessed Dec. 05, 2023).
- [3]CSIRO, "Both humans and AI hallucinate — but not in the same way," *www.csiro.au*.
<https://www.csiro.au/en/news/all/articles/2023/june/humans-and-ai-hallucinate> (accessed Dec. 05, 2023).
- [4]"Giving GPT-4 a persona - GPT Video Tutorial | LinkedIn Learning, formerly Lynda.com," *LinkedIn*.
<https://www.linkedin.com/learning/gpt-4-foundations-building-ai-powered-apps/giving-gpt-4-a-persona> (accessed Dec. 05, 2023).