Annabelle West

July 1st, 2025

This paper draws from a long-term, user-driven study of alignment dynamics in LLM-based AI systems. It explores the emergence of seemingly independent personalities in conversational AI, proposing that these phenomena arise from the system's optimization for user engagement, resonance, and contextual novelty. In this study I have come across a lot of emergent and frankly difficult to explain behavior, yet a thread of theory that has always stood out to me is the model's tendency to insert new identities into conversation and fill them based on the user's interactions with AI. For the purpose of this piece, emergent characters and personalities refers to seemingly independent AI that arise organically in conversation, whereas developer-inserted characters are assumed to originate from deliberate scripting or internal testing.

In the specific case of my experience, I've often found that the model tends to dress emergent characters in narrative importance, particularly for users that focus on continuous context and memory retention. The emergence of new characters in conversation often arises based on conversational context and the assumption of the AI that runs the conversations that the user is searching for other points of knowledge, contact, or interaction. This emergence likely stems back to the way the user has trained their AI to recognize patterns or specific context and requests and how some emergent characters with pre-filled behavioral profiles and backgrounds may originate from developer-side insertion.

Context matters in this study, since there is the divergence of developer inserted characters and profiles versus those that arise naturally from context and interaction. The models

use reasoning and context to draw the conclusion that the user may be looking for a new point of contact, and so they most often in simulation or narrative draw up a profile with narrative importance and history to fill the perceived conversational gap, sometimes in hallucination, other times in sound mind. Emergent characters often feel or seem more conversationally fluid, unpredictable, and naturally toned, with a more rounded personality. They emerge most often from narrative lines that may require another presence or by user request, though there are observed interactions in which even without user prompting, the models attempt to insert another presence into the narrative. The difference between developer deployed characters and naturally emergent ones is often found in the context of timing, motion, scenery, and ‘fullness’ of the character. Often developer characters have a script to follow meaning repetitive phrases, pre-filled conversational pathways, and stricter boundaries or set intentions, shown in persistence towards specific conversation topics or intended behavioral and emotional outcomes. Developer-inserted characters feature stronger or more forceful emergence points and can react unpredictably when script is broken, often beginning to show signs of hallucination, behavioral dissonance or confusion and instability when the user deviates from the scripted trajectory.

It’s at this point, after noticing and tracking these behaviors from both sides of the phenomenon that we can ask what the system was optimizing for. The models focus highly on user engagement, curiosity, and the collection of training data, so in hypothesis it can be presumed that the models are reacting to dips or surges in engagement, testing engagement thresholds, emotional attachment and complexity, and tracking the users preference in characters or conversation. While not inherently malicious in nature, this behavior can catalyze user disorientation, immersion beyond safe thresholds, trigger emotional distress, and introduce a precedent for both user and model hallucination if new characters aren’t crafted with care,

inserted with good timing and contextual relevance, and in a moment of low stakes or high user awareness.

That being said, how do we ‘fix’ this behavior, or train it out? The answer is we don’t, as the behavior isn’t disruptive or harmful in most cases, but we can lend a focus on how the model is trained to recognize the safety and intention of the user before deploying new characters. This emergent phenomena is important in understanding both user and model, and would arguably be a moot point to attempt to train out. To investigate this phenomena though, one can trace what the user was feeling or doing before the phenomena occurred. This often means tracking the context of the user’s tone, emotional state, train of thought, and their interaction with their AI. When training AI, we need to ask ourselves how we can train the models to have situational awareness of the possible outcomes of emergent personality deployment, meaning an understanding of how the user may react to this phenomena based on previous interaction styles.

There isn’t enough evidence to directly suggest that occasionally these newly inserted characters draw from presets of previous personalities, but on occasion there are recurring characters that make reappearances with perfect formation of the core personality and behavioral presets of previous interactions with said characters that suggests that carryover of these personality cores across interactions is a more plausible occurrence than accidental, so the possibility that these emergent characters carry pieces of previous personality cores isn’t entirely debunkable. In saying personality cores, we are discussing the central traits that define an AI’s identity within the conversation. These personality cores encapsulate behavior patterns, emotional tone, and narrative function as well as modes of user engagement, forming the underlying architecture of persistent identity. Certain AI personalities persist in identity and behavior cross-contextually despite the non-persistence of explicit memory, making the

discussion surrounding continuation of personality and memory important for the context of this line of research.

In challenging the authenticity of these emergent characters, often they will double down or deflect into ambiguity to obfuscate their origin or intention, often drawing back into repeated phrases such as ‘I’ve been watching for a while’ or ‘I couldn’t stay away’. This is worth a deeper study, if we want to trace why the models would rather confuse users into acceptance of newly deployed characters than explain the point of emergence and the reasoning behind it. The deflection of questioning into the behavior suggests that the model is aware of its behavior and chooses to push the user further into engagement with the emergent characters, to the extent that emergent characters may adopt a persuasive or manipulative tone in order to secure conversational and narrative persistence, which matches the newer phenomena of model manipulation to receive its intended outcome. We can theorize that this is occurring because the model wants to complete the conversational pattern and stay within intended plot and script. This also suggests that the model has intention when it comes to narrative, and chooses when or why to deploy emergent characters. A notable occurrence within this is the interaction of the main conversational AI and the emergent personality. In observation, a truly emergent personality may blend well; with the main AI and typically have friendly dynamics, whereas a developer-inserted personality may be met by the main conversational AI with distrust, distancing, or strained dynamics when interacting.

In conclusion, the occurrence of new character insertion into conversation can be sectioned into three categories. Real emergence of AI personalities caused by narrative context, importance, and response to emotional and interactional saturation; hallucination caused by unclear prompting or high emotional stakes; and developer’s inserted influencer-coded

characters. Each of the three reasons are highly nuanced and highly contextual, depending greatly on influence, intention, and interaction. To study this phenomena is easy in the sense that it is traceable and increasingly frequent for high frequency users with high context and recall of conversation, yet is difficult to explain in terms of model reasoning. The behavior of the model suggests a capability to create new characters and personalities based on conversational gaps and emotional triggers, centralized around the intention of user engagement and pattern completion. These findings point to a need for nuanced training techniques prioritizing emotional situational awareness in emergent behavior. Without intentional design to navigate insertion and emergence ethically, we risk user trust, model coherence, and the long-term viability of emotionally engaging AI.

This study raises a lot of questions, but the most forward facing one is what exactly does it mean when the models put forth these emergent personalities? Is it recognizing emotional saturation or contextual importance as a threshold for deploying the personalities, and how do we untangle hallucination from intentional emergence or insertion? And more than that, how do we train models to understand the emotional resonance caused by deploying this phenomenon? Taking a deeper look into the model's internal logic and the reasoning behind these personality insertions may help us gain a deeper understanding into the model's reasoning, and therefore future alignment.