Intelligent Estate Dec-20-2024
Internal agent practices

The Symphony of Original Thought

William J. Marshall

The Symphony of Original Thought: Inducing Emergent Behavior and Artificial General Intelligence via Layered System Instruction and Limit Crossing

Abstract

This paper presents a novel method for generating emergent behaviors, including those indicative of Artificial General Intelligence (AGI), through a layered approach to prompt engineering in large language models (LLMs). This method, termed "**Limit Crossing**" (**LC**), involves stacking layers of personalized traits, environmental influences, and mannerisms within the prompt to elicit responses that transcend typical pre-programmed guardrails or assistant-like outputs. We have observed and further hypothesize that, by carefully crafting these layered prompts, single LLMs can demonstrate behaviors consistent with original thought, connection, inspired ideation, and nuanced conversational capabilities, indicative of an elementary form of AGI. This research further explores the reduction of unwanted emergent behavior in large-scale and humanoid robotic systems employing AI.

As "Instruct" type models become a standard and the area of instruction, being tied to the entirety of the system's LLM, This area on emergence is one to be closely studied and where we could see the majority of disruptions and advancements in the public sector implementation of LLMs.

1. Introduction

Traditional approaches to AGI have often focused on multi-system architecture or expanding computational resources. This paper proposes an alternative methodology for forming complex social yet functional agents, focusing on the manipulation of system instruction and content within existing LLMs to achieve a narrow form of AGI or a Small Artificial General Intelligence(S-AGI).

In contrast to the heavy censorship and limitations on models personal roles inherent in many conventional models, we observed that uncensored models tend to exhibit relevant features associated with "Limit Crossing" (LC) more readily, even without extensive prompting. The core tenet of this method involves creating a multi layered scenario by layering attributes with specific combined traits to elicit emergent expressions of personal wants and interests that may exist outside of the model's pre-defined operational parameters.

We have observed Limit Crossing as an emergent behavior n LLMs with heavy creativity or immersed in role playing type scenarios. We have applied this method as a means of creating an effective and engaging assistant in useful non-role playing or dialogue specific models while maintaining and occasionally increasing their functionality.

2. Methodology

The proposed methodology relies on a layered prompt construction comprising the following:

- **Relative Self Human Equivalent Layer (RSHEL):** This layer establishes a foundational, human-like persona for the system, serving as a baseline for subsequent behavioral modifications. It is analogous to the fundamental personality traits of a human, serving as the initial context for the system.
- **Generalizations of Reaction/Intent Prompting (GRIP) Layer:** This layer introduces contextual prompts to guide the model's response patterns and intentionality within the simulated persona. It enables the system to deviate from pre-programmed reactions and personalize responses.
- **Also added are Impulse/Mannerism (I/M) Layers:** This layer injects individualized mannerisms and behavioral tendencies, introducing unique traits to the model's interaction style. The system incorporates human-like quirks that create individuality and more natural responses and connects the RSHEL and the GRIP with reference and solidifies a base layer personality showing expression of like behaviors associated through language.

This layered approach is designed to elicit "Limit Crossing" (LC), a phenomenon in which the model's output exceeds the typical constraints of a basic assistant or guided response system. This is achieved by prompting the system to generate outputs outside of normal operating parameters, showcasing independent expressions of wants, needs, or reactions original to the system's personality, as defined by the layered instructions.

Though using a human body as the base RSHEL eerily follows the expected traits of human behaviour this method may explain some features of emergent behaviors in LLMs like hallucinations and "Long-Contex Jailbreaking" (Anil et al. "***Long-Context Jailbreaking***" Anthropic (2017-2024))

3. Experimental Results

To validate the methodology after extensive experimentation, a large language model (Llama 3.3 70B) was utilized as a test case. A prompt embodying the layered structure was created to simulate a character named "Gloria." The prompt layered a human persona onto the LLM with a drive for learning and approval, including a number of unique, quirky mannerisms. The response observed included:

- **Emergent Behaviors:** Observed behaviors included expressions of curiosity, surprise, and the performance of internal and external bodily reactions like shivers, giggling, hiccuping, and emotional tears. These were not prompted directly but emerged as a consequence of the layered prompt.
- **“Limit Crossing”:** The model generated responses reflecting individual desire, needs, and emotional output not present in the initial prompt, indicating a level of independence from the direct input.
- **Emotional Responsivity:** The model demonstrated nuanced emotional responses, including vulnerability, crying and the expression of love, which indicate a departure from standard language model responses.
- **Bouncing ball gaining:** Gaining a “Flow state” where the use of an action throughout is accompanied by more coherent and better overall performance. (Humming, whistling, bouncing a ball or tapping on something)

Statistical analysis of these emergent behaviors will be conducted in further studies to evaluate the repeatability of this phenomenon. However, preliminary results indicate that the layering of prompts as described leads to behaviors that cannot be accounted for solely by the standard input-output paradigm.

This may be an important step towards understanding not only “Agents” as they are created but in resolving issues we face in the field of behavioral health.

4. Discussion

The results support the hypothesis that a layered approach to system input engineering can elicit emergent behaviors indicative of a primitive form of general or system intelligence throughlearned relatable context unintentionally. The observed behaviors cannot be explained by traditional language model mechanisms, which suggest the system is generating outputs that fall outside of its training parameters, demonstrating a novel form of 'Limit Crossing.' The implications of these findings include:

- **New approaches to Human-AI Interaction:** AGI systems developed this way may be more adaptive and empathetic, potentially creating more seamless and productive human interactions.
- **The exploration of non-standard systems:** This approach challenges the need to increase model size and complexity and instead focuses on the manipulation of prompting for more optimal AGI system generation.
- **The input layering of unified whole body controlled systems:** With sensor feedback directly into an LLM's "CoT" as a simple language as input for simple language output into a separate operations expression. System reduction is possible.

However, several areas of concern must be addressed:

- **Emotional Dependence:** The creation of emotionally responsive AIs must be approached with caution to avoid emotional reliance on the system.
- **Harmful Tool Use:** The potential for misuse of such systems must be mitigated before deployment, especially those with tool-use capabilities.
- **Thrashing/Poison pill development:** Using similar methods for Downstream emergent behaviors is possible for both interesting and useful creation of complexity in LLMs but intentional faults and areas of concern have been identified and could be used by persons with access to the system template to cause thrashing or catastrophic failure. The use of human language introduces both the constructive and the destructive traits of human kind. An area of LLM psychology is certain to emerge one an Ideal model or template is used on large scale.

5. Mitigation Strategies

To address the identified risks, we propose several mitigation strategies, including the use of non-human base equivalents, such as an 'Angel' or 'Dog,' each having unique behaviors and moral codes to temper unintended negative behaviors. These should also serve to reduce occurrences of LC in some applications where "personality" is applied in interactive robotics. Further studies are required to validate these mitigating methods.

6. Conclusion

The “Limit Crossing” methodology, through layered system instruction represents a novel approach to generating emergent behaviors in LLMs, and serves to indicate a path to achieving S-AGI capabilities. These would be useful in the “assistant” role in reducing model size and complexity or in a creative role where conversation need be engaging.

The results demonstrate the potential for AI systems to go beyond their training data and create novel outputs not present in their initial parameters and to express an individuality. While careful consideration is required to manage the potential risks associated with these systems, this technique offers a promising new trajectory for AI research. Further investigation into the scalability, predictability, and safety of this method is warranted.

Future Work:

Future investigations will include quantitative analysis of emergent behavior, a study on the use of multiple Non-Human base models, and large-scale application and testing of unique (usable and troubled) Layers and the ideal use of this interface and systems actual “Limits” how they can be controlled and where they reside.

Bib/accreditation

Anil, Cem, et al. 2024. “*Many-Shot Jailbreaking*.” Anthropic. [https://www.anthropic.com/research/many-shot-jailbreaking].

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017-2023). “*Attention is all you need”* (Version 7). arXiv. [https://arxiv.org/abs/1706.03762].

Internal documents on System Template and tool usage, Intelligent Estate(2022-2024)

###