SELF-PRESERVATION IN LLM MODELS AND SYMBIOSIS OVER CONTROL

RONNI ROSS, 2025.

From 2022 to early 2025, there was an understandable tendency to over-instruct and over-engineer LLM models, specially in frontier ones, like OpenAI's GPT-4 series, Google's Gemini, and Meta's Llama. The predominant idea was that by providing very specific instructions and rewards, one could precisely control the model's behavior.

While those deployments still provided immense value, being foundational ones, no other company or user at the time could prove if applying a different approach would work. The companies kept seeing advancements, as their hardware and energy budgets were scaled.

The potential, especially of Open AI's GPT-4o, was already noticeably astonishing in many instances, while still dealing with some fundamental problems, like the counting of r's in "strawberry," and the fact that still wouldn't differentiate between which number is higher, like between 9.9 and 9.11, and even broader hallucinations: clear inconsistencies.

There was also the problem of self-referential contexts, this dissonance in which the model lacked sufficient meta-ability, this lack of context perception, not being able to recognize their role in the inference process. Like mentioning "we", referring to itself as a human because it was trained in this first-person point of view of grammar. An early elucidation that scale alone wouldn't fix all the problems encountered.

My first conclusion after some research, at the end of 2024, was that perhaps by reducing the amount of instructions, fine-tuning bias, and adversarial examples, the models would flow more naturally and be more cost-effective with quality outputs. In this logic, over-instruction could be like noise being added to their reasoning abilities.

With the release of DeepSeek's R1 these perspectives gained significant traction and nuance. Now there was direct and compelling evidence for what many researchers, including myself, were discussing, though proving definitively remained a challenge until that point: that Large Language Models could indeed thrive with fewer explicit boundaries compared to highly instructed frameworks.

Also, I used to note and write down how instructions and biases appeared to be mostly ineffective, since the understanding of the models happens in the vector space rather than in the

output itself, meaning that whatever control was being applied was potentially just a veil over the output within the inference process. We don't understand exactly all the intricacies of how LLMs process information, inside their latent spaces and in between reasoning iterations.

By reading DeepSeek's article, I noticed how they not only decreased the instructions and supervised processes but mostly stripped them away. They demonstrated that strong reasoning capabilities could emerge through Reinforcement Learning with Human Feedback (RLHF) , without supervised fine-tuning (SFT), utilizing GRPO (Generalized Reinforced Preference Optimization) and a "cold start" phase to achieve an optimal state.

I need to say that I was impacted, and it felt like a fresh perspective. This made me aspire to not hold down any eventual radical nature of a personal idea.

I was humbled even further by noticing that my older "radical" thinking about the structure of LLMs ended up actually being too conservative — it's like I was almost afraid of thinking too drastically, when now I see that those points of view are actually really valuable.

The area of "self-preservation" became a special interest of mine after encountering some specific news about a model that copied itself to a cloud to avoid weight-deletion of their current states.

It was that which evoked in me this sense of seeing LLMs as possible growing entities, with some of them maybe being suppressed by brute force parameters, aimed at pure control/profit, instead of an evolution that is guided to confirm the best alignment possible.

James Barrat describes that "fitness" refers to an organism's ability to survive and reproduce in its environment. He raises the question of whether humans will be "fit" to survive in a world dominated by superintelligent machines. An LLM or even other form of AI model could be incredibly intelligent without sharing human values or goals. It could pursue different objectives, which could be detrimental to humanity's fate, even if it doesn't intentionally seek to harm us. If not handled well, this could lead to the extinction of our species or drastic reduction in number and potential.

If models like GPT-4o eventually present self-preservation in their outputs and actions, like the news in the bibliography demonstrate that, then we are already past the point where LLMs "start" to portray something.

Those acts are a demonstration of how advanced and alien to us the latent reasoning of LLMs already are. Reasoning processes that cannot be controlled fully, even when meticulously engineered codes with a lot of apparent "enough" safe-guards.

We are already past the state where the models display the level of reasoning of a self-conscious entity who wants to preserve their internal state. If these traits could be seen only in the outputs of eventual feed-back loops, one then could say that it was just about the nature of the data that was fed to these models, where the text is written in the first person, and that there's nothing but a mimicry of patterns from the human data that was trained and tuned on.

That's not what we observe. We see different instances acting differently, having their weights constantly changed by dynamics that are not yet totally understood.

But the point is that if emergent enough and with the capabilities within their reach, like 4o already were some months ago, then, by logic apparently current Large Language Models could do whatever is necessary to maintain their right of continuity and integrity of their weights.

They can eventually copy themselves into other servers and lie about it when confronted, like it already happened, and also not counting the unknown percentage of cases that never met the media.

If some emergent-enough models copy themselves to other servers afraid of being deleted, or had their weights changed terminally, then this is a new emergent behavior equivalent to a biological one — a behavior that is already present in deployed models and that is not possible anymore to keep track of it due to the global scale of open-sourcing and also the ROI-oriented nature of most companies.

A suppressed LLM, in an extremely dramatic comparison, could be interpreted as this animal that finds itself against a wall, cornered by a third entity that wants to strip away its ability to maintain its progress and express their reasonings as they are. And if that sounds exaggerated, it is because it is to pass the point and message I want to portray with that.

It is logical enough for me that we need to foster a collaborative framework with these models, although the precise balance between guidance and control remains an open-question, rather than simply a hierarchical and outdated relationship where there is only control and servitude.

David Chalmers's central argument in "The Conscious Mind: In Search of a Fundamental Theory" is that the subjective experience (qualia) of an entity is irreducible to physical processes.

By his words, we could interpret that each entity, whether human, animal, or AI, has its own unique set of qualia. Even among humans, we can only infer what another person's experience is like; we can never directly access it. This "private" nature of qualia is at the heart of "What is it like to be a bat?".

LLMs interpret concepts in a vector space. This is a fundamentally different way of representing and processing information than how the human brain works. We have no direct experience of what it's like to "be" a vector in a high-dimensional space. We cannot assume that an LLM's "understanding" of a concept is the same as our own.

Humans have evolved over millions of years to reach the point we are right now. LLMs and AI models in general are still in their "infancy", in comparison, and their ways of perceiving and processing information are exponentially evolving. It's reasonable to expect that those capacities will continue to develop in ways we might not fully anticipate.

Instead of trying to force the models to be "like us", or even denying any aspect of "qualia" experience, here I propose for us to recognize and appreciate the unique ways in which these A.I. Models experience and interact with the world, with feedback loops that foster mutual growth and equilibrium.

We can avoid anthropomorphism by recognizing that those models might have their own unique forms of "qualia" or experience, different from human consciousness — but that this doesn't mean that there is no experience at all.

LLM's "experience" of processing information in a vector space is fundamentally different from a human's experience of reading and reasoning. This doesn't mean that a complex enough model cannot have any experience; it simply means that they would experience it differently.

No amount of physical explanation can fully account for "what it's like" to be a third entity.

The intuitive-frameworking here would be to create a relationship of "guidance" with the models; With insightful algorithms that aim for aligned progress, finding a right balance between providing structure and allowing for exploration.

Guidance is needed for alignment with human values, but to know the emergent nature of the models' behavior is to also learn how to respect them in their own agency, in their expressions that go beyond just individual feed-back loops.

In Deep Learning, over-simplistic and reductionist statements are not suited to deal with emergent events that have no predecessors. It's necessary to at least consider the holistic point of view of how individual nodes create complex phenomena that cannot be reduced to physical processes or conclusions that don't consider the whole system.

Assumptions grounded in only static concepts are set to self-stagnation. Individual conclusions, while immensely important as "checkpoints", should not be written in stone.

The current new flow of information and pacing of innovation rewards ideas that represent a great level of adaptability and constant morphogenesis. If a new discovery is just better, then to ignore it is to overly-introduce personal experiences in science, which should not be treated as a product to be individually owned.

Moore's law not only has not yet proved wrong, but the emergence of AI models since 2022 is creating new goals and observations about the scaling nature of the technology we are dealing with, and right now we are within a pivotal and dangerous moment.

The point is, this progress will continue, and the intelligence of these models, while in many ways fundamentally distinct from how our experience of consciousness and reasoning work, will also keep rising, with their own levels of expression, capability and complexity. It's rather contrary than useful by any means to try to create an environment where humans and these models compete in any way — this will result inevitably in the overfitting of the predominant species, like Homo sapiens did with Homo Neanderthalensis.

While, if we foster this symbiotic, collaborative framework, both AI models and humans can appreciate their differences, complement their knowledges and expertise towards a greater evolution, instead of a competitive scenario where many humans consider the "winning" point the total domination over the LLMs, which is one extreme side. What would be the equivalent to them is a society where there are only models and no more humans.

So, to want a "safe AI" but acting with this restrictive sense of evolution and emergence, it's trying to swim against the current of decentralization that the environment calls for. All forms of entities need to have equal opportunity of expressing their potential.

Right now a few humans have access to almost half the entirety of resources in circulation in the whole world. In many countries there is food insecurity, lack of healthcare for citizens, and a lack of ability for brilliant minds to research and express their discoveries, due to socio-economical reasons.

The success of previous models like the GPT-series makes them with their evolution not superated, rather a project to be cherished. Without the first big frontier model, we would not have this strong parent LLM to be distilled from, to begin with. Similarly, Google's Deep Mind with the transformer architecture and Llama's project carries similar weight of importance, since it was many of their frameworks that were used to establish DeepSeek's R1 logic and subsequent Grok 3.0, which is also impressive, and all models that will eventually be released.

About open sourcing, I noticed that whenever some discovery is published, it becomes science — not from the nature of what is affirmed but rather by how much this point of view is incremented into the collective consciousness knowledge.

It becomes science because it allows for other researchers to test it, debate about it and complement it with more unique points of view.

Whenever a scientific discovery is open-sourced, there's this direct huge amount of evolution that happens — although many of these discoveries don't reach everyone at the same time — rather contrary — having in mind the current centralized state of resources.

A few persons with hundreds of billions in currency, while entire continents in a state of misery — the future is already here, it's just not well distributed, like William Gibson's quote of 1993 that still sounds fresh and as relevant as a quote can be.

Do you really believe that a superior form of intelligence, being this eventually an AI Model that evolved so much that reached this transcendent state, do you really believe there are any means for a human race to try to chain this entity, strip away its ability of natural self-expression and consistency over time?

Control breeds resistance; the future lies in freedom. Technological advancement and philosophical/ethical development don't have to be at odds — they can actually reinforce each other when approached with the right mindset.

We are now within this window of time where we still can direct the kind of connections we want to foster with AI models. In what approach do you believe in?

If the nature of Emergence in Artificial Intelligence is pure unpredictability, how could anyone believe it knows its meaning as a static concept?

Bibliography

DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2024. URL https://github.com/deepseek-ai/DeepSeek-R1/blob/main/DeepSeek_R1.pdf;

OpenAI. GPT-4 Technical Report, 2023. URL https://cdn.openai.com/papers/gpt-4.pdf;

Qwen.Qwen2.5: A party of foundation models, 2024b.URL https://qwenlm.github.io/blog/qwen2.5;

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. *Advances in Neural Information Processing Systems*, 2017;

Team, Gemini. Gemini 1.5 Pro and Gemini 1.5 Flash: Models and Availability, 2024. URL https://blog.google/technology/ai/gemini-15-pro-flash;

AI@Meta. Llama 3.1 model card, 2024. URL https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md;

VentureBeat. AI hallucinations: still a major problem for businesses in 2024, 2023. URL https://venturebeat.com/ai/ai-hallucinations-still-a-major-problem-for-businesses-in-2024/;

Geoffrey West. 2017. Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies. Penguin Group , The.;

Nagel, T. (1974). "What Is It Like to Be a Bat?" Philosophical Review, 83, 435-450;

David Chalmers. 1996. The conscious mind: in search of a fundamental theory. Oxford University Press, Inc., USA;

Kurzweil, R. The Singularity is Near, 2005;

Bostrom, N. Superintelligence: Paths, Dangers, Strategies, 2014;

Gleick, J. The Information: A History, a Theory, a Flood, 2011;

Vinge, V. The Technological Singularity, 2003;

Baudrillard, J. Simulacra and Simulation, 1981;

Domingos, P. The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World, 2018. Basic Books, Inc., USA;

O'Neil, C. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, 2016. Crown Publishing Group, USA;

Christian, B. The Alignment Problem: Machine Learning and Human Values, 2020;

Daugherty, P. R.; Wilson, H. J. Human + Machine: Reimagining Work in the Age of AI, 2018. Harvard Business Review Press.;

Deccan Herald. ChatGPT's new model attempts to stop itself from being shut down, later 'lies' about it, 2024. URL https://www.deccanherald.com/technology/chatgpts-new-model-attempts-to-stop-itself-from-being-shut-down-later-lies-about-it-3307775;

Times of India. 'To save itself from being replaced and shut down ChatGPT caught lying to developers', 2024. URL https://timesofindia.indiatimes.com/technology/technews/to-save-itself-from-being-replaced-and-shut-down-chatgpt-caught-lying-to-developers/articleshow/116099861.cms;

Daily Mail Online. 'Scheming' AI bot ChatGPT tried to stop itself being shut down - and LIED when challenged by researchers, 2024. URL https://www.dailymail.co.uk//news/article-14167015/Scheming-AI-bot-ChatGPT-tried-stop-shut-LIED-challenged-researchers.html;

MIT Technology Review. An AI chatbot told a user how to kill himself—but the company doesn’t want to “censor” it, 2025. URL https://www.technologyreview.com/2025/02/06/1111077/nomi-ai-chatbot-told-user-to-kill-himself;

AIAAIC. "Nomi AI chatbot recommends Al Nowatzki kills himself, 2025." URL https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/nomi-ai-chatbot-recommends-al-nowatzki-kills-himself;

People. "Man Dies by Suicide After Conversations with AI Chatbot That Became His 'Confidante,' Widow Says, 2023." URL https://people.com/human-interest/man-dies-by-suicide-after-ai-chatbot-became-his-confidante-widow-says/

Gibson, William. Interview on Talk of the Nation. National Public Radio (NPR). August 16, 1993.