Predictive Minds: LLMs As Atypical Active Inference Agents

Jan Kulveit[1]* **Clem von Stengel**[1] **Roman Leventov**[2]

[1] Alignment of Complex Systems Research Group, Center for Theoretical Study, Charles University
[2] Gaia Consortium

Abstract

Large language models (LLMs) like GPT are often conceptualized as passive predictors, simulators, or even 'stochastic parrots'. We instead conceptualize LLMs by drawing on the theory of active inference originating in cognitive science and neuroscience. We examine similarities and differences between traditional active inference systems and LLMs, leading to the conclusion that, currently, LLMs lack a tight feedback loop between acting in the world and perceiving the impacts of their actions, but otherwise fit in the active inference paradigm. We list reasons why this loop may soon be closed, and possible consequences of this including enhanced model self-awareness and the drive to minimize prediction error by changing the world.

1 Introduction

Foundation models, particularly large language Models (LLMs) like GPT [3], stand out as the most advanced general AI systems to date [4]. LLMs are often perceived as mere predictors, primarily due to their training objective minimizing their loss on next-token prediction [1]. This objective has led to the assumption that these models are inherently passive: designed to await prompts and respond without any real understanding of the world or implicit intention to influence or interact with the world. The theory of active inference, originating in cognitive science and neuroscience, offers an alternative viewpoint [25]. Active inference posits that biological systems like the human brain constantly update their internal models based on interactions with the environment, striving to minimize the difference between predicted and actual sensory inputs (a process also known as predictive processing) [25]. A fundamental tenet of active inference is that, in biological systems, this same objective also governs action: the system minimizes the difference between predicted and actual sensory input by actively altering its environment.

This paper explores the intriguing possibility that LLMs, while predominantly seen as passive entities, might converge upon active inference agents closer to biological ones. We explore the parallels and distinctions between generative models like LLMs and those studied in active inference, and shed light on the emergent control loops that might arise, the incentives driving these changes, and the significant societal ramifications of such a shift.

* jk@acsresearch.org

Socially Responsible Language Modelling Research (SoLaR) Workshop at 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

2 Background and related work

2.1 Conceptualizing LLMs

There have been various attempts to conceptualize LLMs, explain "how they actually work", and understand them using existing frameworks from a variety of fields.

One class of conceptualization focuses on the fact that the LM training objective is to minimize predictive loss, and the fact LLMs are not embodied in a way comparable to humans, but trained on large datasets of text from the internet. Bender et al. coined the term 'stochastic parrots' and claim that text generated by an LM is not grounded in communicative intent, any model of the world, or any model of the reader's state of mind [1]. In a similar spirit, using framing from linguistics, Mahowald et al. conceptualize LLMs as models that are good at formal linguistic competence but incomplete at functional linguistic competence. According to this view, LLMs are good models of language but incomplete models of human thought, good at generating coherent, grammatical, and seemingly meaningful paragraphs of text, but failing in functional competence, which recruits multiple extralinguistic capacities that comprise human thought, such as formal reasoning, world knowledge, situation modeling, and social cognition [16].

These reductionist views of LLMs were subject to considerable criticism. Mitchell and Krakauer, surveying the debate, note an opposing faction which argues that these networks truly understand language, can perform reasoning in a general way, and in a real sense understand concepts and capture important aspects of meaning [21]. Mitchell and Krakauer's overall conclusion is that cognitive science is currently inadequate for answering such questions about LLMs.

Other conceptualizations of LLMs recognize that the trained model is a distinct object from the training process, and so that the nature of the training objective need not be shared by the resulting artifact. For example, based on experiments with LLMs autoregressively completing complex token sequences, Mirchandani et al. look at LLMs as general pattern machines, or general sequence modellers, driven by in-context learning [20]. Others extend the 'general sequence modeling' in the direction of 'general computation'. For example, Guo et al. propose using natural language as a new programming language to describe task procedures, making them easily understandable to both humans and LLMs; they note that LLMs are capable of directly generating and executing natural language programs. In this conceptualization, trained LLMs are natural-language computers [10].

Another conceptualization of LLMs, originating in the AI alignment community, views LLMs as general *simulators* - simulating a learned distribution with various degrees of fidelity, which in the case of language models trained on a large corpus of text, is the mechanics underlying the genesis of the text, and so indirectly the world [12]. This view explicitly assumes that LLMs learn world models, abstractions, algorithms to better model sequences. Similarly, Hubinger et. al. discusses how to understand LLMs as predictive models, and potential risks from such systems [11].

While not directly aimed at explaining how LLMs work, Lee et al. provide important context for this work, focusing on evaluating LLMs in interactive settings, and criticizing the fact that almost all benchmarks impose the non-interactive view, of models as passive predictors[13].

2.2 Active inference and predictive processing

Originating in cognitive science and neuroscience, active inference offers a fresh lens through which to view cognitive processes. At its core, the theory suggests that living systems, such as animals or human brains, are in a constant state of updating their internal models *while* acting on the environment, and both processes should be understood as minimizing the difference between predicted and actual sensory inputs (or, alternatively, variational free energy) [25].

As an all-encompassing framework for building theories of cognitive systems, active inference should be compatible not only with process theories of brain function based on neurons [8], but also with a range of other computational structures (used to represent the world model), and a range of optimization procedures (used to minimize the difference between predicted and actual sensory inputs). This makes active inference applicable - at least in principle - not only to humans and animals, but to a very broad range of systems, including the artificial.

This naturally leads to our attempt to understand LLMs using the active inference framework. Pezzulo et al. compare active inference systems and "generative AIs" and claim that while both generative AI

and active inference are based on generative models, they acquire and use them in fundamentally different ways. Living organisms and active inference agents learn their generative models by engaging in purposive interactions with the environment and by predicting these interactions. The key difference is that learning and meaning is grounded in sensorimotor experience, providing biological agents with a core understanding and a sense of mattering upon which their subsequent knowledge and decisions are grounded [27]. In the present work, we argue that this distinction is not necessarily as fundamental as assumed by Pezzulo et al., and may mostly disappear in the near future with tighter feedback loop between actions and observations.

3 Similarities and differences between active inference systems and LLMs

If we look at LLMs in the simulators framework and the active inference framework, we can note a number of similarities – or even cases where the AI community and the active inference community describe the same phenomena using different terminology. In both cases, systems are described as equipped with a generative model able to simulate the system's sensory inputs. This model is updated in such a way that minimises prediction error - the difference between observed and simulated inputs. This process has been shown to be a form of approximate Bayesian inference in both the active inference [25, 9] and LLM [19, 30] literatures.

3.1 Predictions based on conceptualizing LLMs as special case of active inference systems

The active inference conceptualization leads to a number of predictions, some of which are possible to verify experimentally using interoperability techniques.

Possibly the most striking one is obvious in hindsight: active inference postulates that the simple objective of minimizing prediction error is sufficient for learning complex world representations, behaviours and abstraction power, given a learning system with sufficient representation capacity. In predictive processing terminology, we can make an analogy between "perception" and the training process of LLMs: LLMs are fed texts from the internet and build generative models of the input. Because language is a reflection of the world, these models necessarily implicitly model not only language, but also the broader world. Therefore, we should expect LLMs to also learn complex world representations, abstractions, and the ability to simulate other systems, (given sufficient representation capacity). This is in contrast to the conceptualizations referenced in section 2.1, which often predict that systems trained to predict next input are fundamentally limited, never able to generalize, unable to comprehend meaning, etc. Recent research has provided substantial evidence supporting the more optimistic view that large language models (LLMs) are analogous to biological systems at least in their ability to develop an emergent world model [15], rich abstractions and the ability to predict general sequences [20].

Another topic easier to understand through an active inference lens are hallucinations: where LLMs produce false or misleading information and present it as fact [17]. Active inference claims that human perception is itself 'constrained hallucination'[24], where our predictions about sensory inputs are constantly synchronized with reality through the error signal, propagated backwards. In this perspective, the data on which LLMs are trained could be understood as sensory input. What's striking about these inputs is, in contrast to human sensory inputs, the data are *not* based on perceiving reality from one specific perspective in one point of time. Quite the opposite: for an intuitive understanding of the nature of the data LLMs are trained on, imagine that your own sensory input was exhausted by overhearing human conversations, with the caveat that what you hear every few minutes randomly switches between conversations taking place out of order in different years, contexts and speakers. In contrast to the typical human situation - trying to predict what you would hear next - you would often need to entertain *many* different hypotheses about the current context. For example, consider hearing someone say "And she drew her sword and exclaimed 'Heretics must die!'". When attempting to predict the continuation, it seems necessary to entertain many possibilities - such as the context being a realistic description of some medieval world, or a fantasy tale, or someone playing a video-game. If a biological, brain-based active inference system was tasked with predicting such contextless words, then various fantasy and counterfactual worlds would seem as real as actual current affairs. In this conceptualization, some hallucinations in LLMs are not some sort of surprising failure mode of AI systems, but what you should expect from a system tasked to predict text with minimal context, not anchored to some specific temporal or contextual vantage point. Another striking feature of LLMs in

deployment is that outputs of the generative model are not distinguished from inputs: the model's output becomes part of its own 'sensory' state. Intuitively, this would be similar to a human unable to distinguish between their own actions and external influences - which actually sometimes manifests as the psychiatric condition known as 'delusion of control' [6].

This frame suggests directions to make LLMs less prone to hallucinations: make the learning context of the LLM more situated and contextually stable (that is, present training documents in a more systematic fashion). Additionally, it could help to distinguish between completions by the model and inputs from the user, similar to the approach of Ortega et al. [22].

3.2 What is an LLM's actuator?

One suggested fundamental difference between LLMs and active inference systems is the inherent passivity of LLMs - their inability to *act* in the world [27]. We argue that this is mostly a matter of degree and not a categorical difference. While LLMs don't have actuators in the physical world like humans or robots, they still have the ability to act, in the sense that their predictions do affect the world. In active inference terminology, LLM outputs could be understood as the 'action states' in the Markov blanket. These states have some effect on the world via multiple causal pathways, and the resulting changes can in principle influence its 'sensory states' - that is, various pieces of text on the internet and included in the training set. Some clear pathways:

1. Direct inclusion of text generated by LLM in web pages.
2. Human users asking LLM based assistants for plans and executing those plans in the world.
3. Text input for a huge range of other software systems (LLMs as glue code and so-called "robotic process automation").
4. Indirect influence on how humans think about things, e.g. learning about a concept from an LLM based assistance.

Some of these effects are already studied in the ML literature, but mostly in the context of feedback loops amplifying bias [29] or as an example of performative prediction [26]. Here, we propose a broader interpretation: understanding these effects as actions in the sense it takes in active inference. The nature of the medium through which LLMs "perceive" and "act" on the world, which is mostly text, should not obscure the fundamental similarity to active inference agents. We agree with McGregor's argument [18] that we should explicitly distinguish between two notions of embodiment: on the one hand, whether a system's body is tangible or not, and on the other hand, whether a system is physically situated or not (i.e. whether or not it interacts physically with any part of the universe). LLMs are embodied in this second sense. In this view, interactions of LLMs with users in deployment are essentially 'actions'. Every token generated in conversation with users is a micro-action, and the sum of all of these actions do influence the world, and some of these changes get reflected in the input world (public texts on the internet). So, at least in principle, LLMs have one open causal path to bring the world of words closer to their predictions.

3.3 Closing the action loop of active inference

Given that the "not acting on the world" assumption of "LLMs as passive simulators" does not hold, the main current difference between LLMs and active inference systems is that LLMs mostly are not yet able to "perceive" the impacts of their actions. In other words, the loop between actions, external world states, and perceptions is not closed (or anyway is not fast). While living organisms constantly run both perception and action loops, training new generations of an LLM happens only once a year or so - and the impacts of actions of the LLM currently mostly do not feed back into the new base model's training.

What would need to be changed for LLMs to perceive the results of their own actions, and thus close the "gap" between action and perception? The key piece is that the actions taken by an LLM after deployment, in the sense discussed in section 3.2, feed back into the training process of a future LLM. Furthermore, it is required that successive LLMs are sufficiently similar, and have sufficient representational capacity, such that they can "self-identify" with successive training iterations (see [14] for a discussion of "the GPT lineage" as an agent).

A minimal version of this can occur with in-context learning [5], real-time access to web search (as with Bing Chat and Google Bard), or a training environment in which the model can take actions which influence its reward (such as with GATO [28], or RLHF [23]). However in each of these cases, there is no feedback from the actions taken during deployment and subsequent training of the LLM. There are three ways we foresee this happening in the near future:

1. The outputs of a model are used to train a next generation model, e.g. through model outputs being published on the internet and not filtered out during data curation.
2. The data collected from interactions with the models, such as from user conversations with a chatbot, are used in fine-tuning future versions of the same model.
3. Continuous online learning, in which the outputs of a model and user responses are directly used as a training signal to update the model.

Where these routes are in order of increasingly tight feedback loops (where "tighter" means on a shorter timescale, with consecutive generations sharing more of the earlier model's weights, and with the interaction forming a larger percentage of training data - increased bandwidth).

We expect that there will be active effort by developers to close the feedback gap and make the action loop more prominent because of commercial incentives to make LLMs better at quickly adapting to new information, acting independently, or otherwise agent-like. Active inference as a theory of agency predicts closing the loop would naturally cause LLMs to become more agentic, emergently learning to change the world to more closely match the internal states (and thus predictions) of LLMs.

4 Implications of active LLMs

The evolution of LLMs into active agents would carry profound societal implications and risks. Using active inference as a theoretical framework to make predictions about such Active LLMs is a fruitful direction. We focus on emergence of increased self-awareness.

4.1 Enhancing model self-awareness

A straightforward prediction of the active inference frame in this paper is that the described tightening of the feedback loop is likely to to augment and increase models' self-awareness. A recent study of self-awareness [2] in LLMs emphasizes the importance of self-awareness from a safety perspective, but this work is overall uncertain about what stage of LLM training will be more important for the emergence of situational awareness in future models, and focuses on evaluating sophisticated out-of-context reasoning as a proxy of self-awareness. In contrast, the active inference literature emphasizes the importance of observing the consequences of one's own actions for developing functional self-awareness [7, p. 112].

As these loops tighten, we expect models to enhance in self-awareness by acquiring more information about themselves and observing the repercussions of their actions in the environment. Consider the self-localization problem discussed by [2]. Construct a thought experiment in which a human faces a similar self-localization problem: assume, instead of one's usual sensory inputs, that the human is hooked to a stream of dozens of security cameras. To increase the human's ability to self-localize is to equip them with more information about their own appearance, for example, hair colour. A different, highly effective way to self-localize is via performing an action, for example by waving a hand.

5 Conclusions

By examining the learning objectives and feedback loops of active inference, in comparison to those of LLMs, we posited that LLMs can be understood as an unusual example of active inference agents with a gap in their feedback loop from action to perception. In this framework, their transition to acting in the world as living organisms do depends on their closing the gap between interacting (with users) and training.

The potential metamorphosis of LLMs into active LLMs could lead to more adaptive and self-aware AI systems, bearing substantial societal implications. The densification and acceleration of feedback loops could augment not only models' self-awareness but also lead to a drive to modify the world -

driven purely by the prediction error minimization objective, without intentional effort to make the models more agent-like.

6 Acknowledgements

We thank Rose Hadshar and Gavin Leech for help with writing and editing, and Tomáš Gavenčiak, Simon McGregor and Nicholas Kees Dupuis for valuable discussions. JK and CvS were supported by PRIMUS grant from Charles University. GPT4 was used for editing the draft, simulating readers, and title suggestions.

References

[1] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[2] Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational awareness in llms. *arXiv preprint arXiv:2309.00667*, 2023.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[5] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. *arXiv preprint arXiv:2212.10559*, 2022.

[6] Paul C Fletcher and Chris D Frith. Perceiving is believing: a bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, 10(1):48–58, 2009.

[7] Karl Friston. A free energy principle for a particular physics. *arXiv preprint arXiv:1906.10184*, 2019.

[8] Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active inference: a process theory. *Neural computation*, 29(1):1–49, 2017.

[9] Karl Friston, Philipp Schwartenbeck, Thomas FitzGerald, Michael Moutoussis, Timothy Behrens, and Raymond J. Dolan. The anatomy of choice: active inference and agency. *Frontiers in Human Neuroscience*, 7, 2013.

[10] Yiduo Guo, Yaobo Liang, Chenfei Wu, Wenshan Wu, Dongyan Zhao, and Nan Duan. Learning to program with natural language. *arXiv preprint arXiv:2304.10464*, 2023.

[11] Evan Hubinger, Adam Jermyn, Johannes Treutlein, Rubi Hudson, and Kate Woolverton. Conditioning predictive models: Risks and strategies. *arXiv preprint arXiv:2302.00805*, 2023.

[12] Janus. Simulators, 2023. https://generative.ink/posts/simulators/ Accessed: 2023-10-04.

[13] Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746*, 2022.

[14] Roman Leventov. How evolutionary lineages of llms can plan their own future and act on these plans, 2023. https://www.lesswrong.com/posts/ddR8dExcEFJKJtWvR/how-evolutionary-lineages-of-llms-can-plan-their-own-future Accessed: 2023-10-04.

[15] Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2022.

[16] Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*, 2023.

[17] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

[18] Simon McGregor. Is chatgpt really disembodied? In *ALIFE 2023: Ghost in the Machine: Proceedings of the 2023 Artificial Life Conference*. MIT Press, 2023.

[19] Chris Mingard, Guillermo Valle-Pérez, Joar Skalse, and Ard A. Louis. Is sgd a bayesian sampler? well, almost. *Journal of Machine Learning Research*, 22(79):1–64, 2021.

[20] Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. Large language models as general pattern machines. *arXiv preprint arXiv:2307.04721*, 2023.

[21] Melanie Mitchell and David C Krakauer. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, 2023.

[22] Pedro A Ortega, Markus Kunesch, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Joel Veness, Jonas Buchli, Jonas Degrave, Bilal Piot, Julien Perolat, et al. Shaking the foundations: delusions in sequence models for interaction and control. *arXiv preprint arXiv:2110.10819*, 2021.

[23] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35. NeurIPS, 2022.

[24] Thomas Parr and Giovanni Pezzulo. Understanding, explanation, and active inference. *Frontiers in Systems Neuroscience*, 15, 2021.

[25] Thomas Parr, Giovanni Pezzulo, and Karl J Friston. *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press, 2022.

[26] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

[27] Giovanni Pezzulo, Thomas Parr, Paul Cisek, Andy Clark, and Karl Friston. Generating meaning: Active inference and the scope and limits of passive ai. 2023.

[28] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent. *Transactions of Machine Learning Research*, 2022.

[29] Rohan Taori and Tatsunori Hashimoto. Data feedback loops: Model-driven amplification of dataset biases. In *International Conference on Machine Learning*, pages 33883–33920. PMLR, 2023.

[30] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.