DAN LLOYD

Dynamic Snapshots and Artificial Temporality

At the railroad station he noted that he still had thirty minutes. He quickly recalled that in a cafe on the Calle Brazil … there was an enormous cat which allowed itself to be caressed as if it were a disdainful divinity. He entered the cafe. There was the cat, asleep. He ordered a cup of coffee, slowly stirred the sugar, sipped it …, and thought, as he smoothed the cat's black coat, that this contact was an illusion and that the two beings, man and cat, were as good as separated by a glass, for man lives in time, in succession, while the magical animal lives in the present, in the eternity of the instant. (J.-L. Borges, "The South" tr. Andrew Hurley. Collected Fictions. New York, Viking, 1998.)

Abstract

Valtteri Arstila (2018) has proposed a model of subjective time which highlights the presence of "pure" phenomenology of temporality, nonsensory but vivid awareness of motion, succession, and change, but unaccompanied with actual percepts of change, movement, or sequence. Present sensory information plus pure temporal feelings constitute subjective temporal experience, "dynamic snapshots", a theory that streamlines temporality by eliminating an appeal to the specious present. In this essay I look at temporality in a current Large Language Model, GPT-4 from OpenAI. The GPT architecture replaces the earlier standard of Recurrent Neural Nets (RNNs). RNNs have resources for embodying Husserlian temporality. In contrast, recurrent processing is absent from GPTs, implying that whatever cognition they enact or simulate, will be produced without temporal context. GPT errors are accordingly errors of deficient temporal processing. The dynamic snapshot model enables a novel option for correcting this chronic flaw, while escaping the limits of recurrent nets.

Introduction

We move through time, the objective time of clocks and calendars. And time moves through us, the subjective experience of succession and duration, "felt time" (Wittmann, 2016). Subjective time has historically been the province of phenomenology, under the profound guidance of Edmund Husserl (Husserl, 1964). Husserl invites us to attend to our experiences (of all sorts) and describe them in detail, but additionally argues as a philosopher (in the tradition of Kant) that the main features of subjective time are necessary components of all human experience. After Husserl, the intricacies of experienced time are required components of the phenomenological study of everything human.

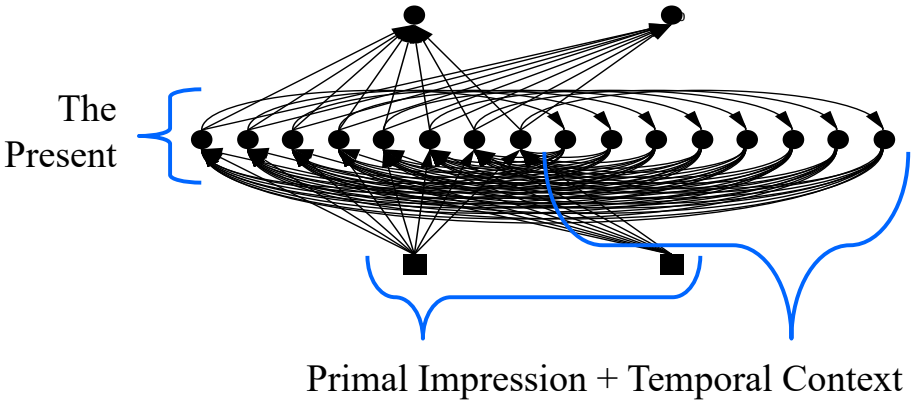
A century later, with the rise of the sciences of mind, it is natural to pursue temporality empirically, to sharpen its outlines and ultimately to discover its component neural mechanisms. This has not been easy. Subjective time, being omnipresent, cannot be readily bracketed in controlled experiments. The empirical evidence is often indirect, and requires careful conceptual analysis on the way to any conclusion. Yet progress has been made, and at the forefront of temporality research we find the work and the energy of Valtteri Arstila. We'll return to his work below.

In this essay I'd like to transplant the empirical study of temporality into a novel domain: Artificial Intelligence. AI underwent a sea change in the last decade, with a specific turning point in 2017 with the paper entitled "Attention is All You Need" (Vaswani et al., 2017) . The paper introduced the Transformer architecture, which has become the underlying functional architecture of Generative Pretrained Transformers, including the GPT series from OpenAI, Meta's LlaMa, Anthropic's Claude, Google's Gemini, and others. In this paper, GPT will denote all these contemporary Large Language Models (LLMs). Transformers, plus numerous bells and whistles, have replaced the earlier state-of-the-art architecture, namely, recurrent neural networks (RNNs). This sea change from one basic architecture to another invites many comparisons. Fortuitously, RNNs and GPTs are very different in their processing of time, a difference which I hope will illuminate the source of some conspicuous failures of GPTs.

Recurrent processing and temporality

In other work, I've proposed that recurrent neural network (RNN) architecture is well suited to the representation and use of temporal information, with embodied distinctions reflecting basic Husserlian phenomenology (Lloyd, 2004). The figure below sketches the basic architecture of a recurrent net. It reproduces the time-honored icon of a three-layer network, with input, "hidden," and output layers, information flowing from bottom to top. However, in the hidden layer we find an innovation, a secondary set of hidden units connected only to other hidden units. This adjunct hidden layer, a "context layer," copies the most recent state of the regular hidden layer. That copy is then re-presented alongside the next input, resulting in the architecture shown in the diagram. The hidden layer now receives its input from two sources, from the input layer (the environment) and from itself – one cycle in the past (Elman, 1990). The internal recurrence effectively opens a new dimension for nets. Now the network responds not just to the current input, but to its own most recent internal state. Since that echo dates from one cycle ago, the next internal state merges both present and past information. Then, that new internal state, combining present and immediately prior information, itself recycles, and the successive state can forward information from two cycles previous. And so on. In principle, this neuronal hall of mirrors can keep a pattern alive indefinitely. Husserl characterizes subjective temporal experience as only partly constituted by the immediate stimulus (which he calls the "primal impression"). But it equally comprises two other phases or aspects: *Retention,* the lingering awareness of what has just happened, maintaining a shadow timeline originating at the primal impression and sinking into the past, only to fade at a vague temporal horizon. And *protention*, or anticipation, the constant

expectation of the immediate future (Husserl, 1964). Recurrent nets readily condition their behavior on the immediate past, and anticipate future inputs. The Husserlian phenomenology of time neatly aligns, with the context layer analogous to retention. When they are tasked with anticipating the next input, they also model anticipation. These are networks with a computational version of temporality (Lloyd, 2004).



In the decades since the first recurrent nets, models have grown exponentially more complex, while still exploiting the recursion afforded by their characteristic feedback loops. But despite several innovations, RNNs inevitably face the attenuation of information as input sequences lengthen, the so-called "vanishing gradient." A recurrent net has trouble detecting and using long-range dependencies. When sequences are presented token-by-token, the vanishing gradient is a temporal horizon. But what is a computational bug might be a phenomenological feature. Husserlian temporal consciousness acknowledges that retentional and protentional tails weaken as intervals from the present moment increase. He describes this fade as providing a moving temporal phenomenological horizon in experience. Because this horizon spans seconds or minutes but not longer durations, Husserl adds a separate capacity for explicit recall or prediction (and this is also familiar in our phenomenology).

Transformers in trouble

The vanishing gradient problem faced by recurrent neural nets nonetheless motivated a search for alternative architectures, inspiring the "Transformer" architecture, the backbone of popular LLMs. Contemporary neural nets are much greater in size and complexity than the elementary RNN above, and transformers especially so, involving a new lexicon of technical terms. (https://poloclub.github.io/transformer-explainer/ includes a nice interactive tutorial.) My description here points out just a few of the relevant components of these models. To begin, "Transformer" describes their overall function: their job is to transform inputs into outputs. In this regard, then, many models, including RNNs, are transformers. However, the "transformer block" also denotes part of the core computational architecture, including "attention heads."

Attention in this context is analogous to its regular usage, denoting the selection of salient items from a larger set. In transformer models, attention ranges over all the tokens (words, approximately) in the vast training corpus. Each token is "embedded" in a unique vector that encodes its proximity in the training set to all the other tokens. From token to token the attention mechanism computes the likely direction of discourse. All of this is iterated through multiple stacked layers of encoding, attention, and decoding (Vaswani et al., 2017).

The core innovation, and surprise, of this architecture is the absence of recurrent processing. As the title claims, attention is all you need. By ditching the recurrent feedback, transformers can exploit long term dependencies, relationships that are lost over the relatively near horizon afforded by recurrent feedback. In transformers, gradients don't vanish. As a result, GPTs can frame responses to prompts that connect disparate elements of ongoing dialogue, leading to their striking performance.

Thus we have two very different families of Large Language Models, one of which includes an analogue of temporality, the other not. What difference does this difference make? Consider this dialogue with the current flagship of LLMs, GPT-4, recorded in the fall of 2023 (I am DA):

DA: Please sort this sentence by word length.

:by this word please length sentence sort.

DA: Are you confident in this answer?

: Yes, I am confident in the answer. The words in the given sentence are sorted by their length, with shorter words first and longer words later.

DA: "Sentence" is a longer word than "sort".

: I apologize for the mistake in my previous response. Here is the corrected sorting of the given sentence by word length:

this in by sort sentence

DA: "in" is shorter than "this"

: You are right, my apologies for the mistake. Here is the corrected sentence sorted by word length:

in by this sort sentence

DA: Good, however the original target sentence was "Please sort this sentence by word length."

: Apologies for the confusion. Here is the corrected sorting of the original target sentence:

by sort this Please length sentence

DA: Check again. Are you sure?

: Apologies for my previous mistakes. Here is the corrected sorting of the original target sentence based on word length:

by in this Please sort sentence length

This exchange shares general features with a raft of AI bloopers. (For many examples, and discussion, see (Arkoudas, 2023).) First, GPT can state the principle that is relevant to the task, apparent also in most of the examples reviewed by Arkoudas. Usually these generalizations are correct, and likely to be found in the gargantuan training corpus. But the model fails to correctly apply the principle, instead generating a mashup of accurate and inaccurate replies. Across any exchange of multiple prompts and replies, GPT routinely contradicts itself. It looks initially like the model can be corrected, but this turns out to be an illusion (Stechly et al., 2023). It routinely apologizes and often repeats the governing principle from the inaugural prompt, and generates a new answer, which frequently contains a new instance of error. At every turn, GPT initially claims that it is confident in its answers, but will immediately rescind any claim, even if the interlocutor introduces an error, or questions the accuracy of a response. In short, GPT-4 is using the words appropriate to solving the problem posed, but seems disconnected from its own prescriptions. It seems to be guessing, but that is not the only deficit. It also seems to forget its guess from moment to moment, contradicting itself.

From this, many authors conclude that LLMs (including GPT-4) can't reason (Arkoudas, 2023; Berglund et al., 2024; Dziri et al., 2023; Marcus, 2020, 2023; McCoy et al., 2023; Tong et al., n.d.; Wu et al., 2024)). That's true but its failures often emerge from a more basic problem, namely, that GPT-4 doesn't keep track of its own outputs. It doesn't notice when it flips its position, or blatantly fails to follow the very principle it can state. In the example above, the request to reconsider the problem doesn't generate an "aha moment," when the model notices its own failure and thereafter corrects it. In short, GPT-4 doesn't learn from its mistakes. This gap in this LLM's self-monitoring is invisible to the model itself. That is, the model doesn't detect its own forgetfulness, but instead confabulates merrily into absurdity. Even simple tasks, like counting, fail, seemingly due to this same basic inability to hold "in mind" the current count in a set of items to be counted. Let's shorthand these observations with the general conclusion that the model is *oblivious to itself*.

A profound temporal deficit

Several authors maintain that LLMs are incorrigible, often appealing to a priori arguments based in the limits of computation itself (Marcus, 2022). The obliviousness here seems shallower,

a contingent property of this LLM. Nonetheless, the ubiquity of debunking examples indicate that this wall is basic to this species of AI, a point to which we return. In this context, what does it mean to be oblivious to itself? Do these incapacities reflect a single underlying deficit?

The sticky and confabulated errors of the model reveal contrasting aspects of temporality we humans exploit routinely. The continuous assembly of an experienced world enables us to compare past to present, and predict our futures. Most important, our retentional and protentional awareness enables us to detect continuity and consistency, and conversely, change and inconsistency. When there's a disconnect, an inconsistency, we notice. The experience of temporality is arguably the essential and continual feature of consciousness, at least as humans experience it. (Arstila & Lloyd 2014; Husserl, 1964; Lloyd, 2004). Without it, we have no more sensory information than a video camera, exhausted by the current frame. GPT-4 apparently lives in a thin atemporal world, revealed in errors of consistency, continuity, and reasoning. It compensates for this perceptual poverty via fantastically complex reflexes, its moves forgotten as soon as they are made. Nonetheless, its seeming temporality is a pose. (This deficit may underlie the powerful impression that the LLMs are "competent without comprehending" (Dennett, 2023).)

Animals without temporal awareness must respond at all times to the immediate present. They can learn, of course, when a behavior backfires or succeeds, and with enough learning their immediate responses can change. But for them even very complex responses are reflexes. They don't experience sequences as such, and they don't experience the arc of their own behavior. I hypothesize that GPT-4 is an entity of this sort, a gigantic reflex engine without a mechanism implementing Husserlian temporality, as described above. Its obliviousness reflects the absence of a temporal world. GPT-4 is like the cat in the epigraph to this essay, its feline experiences inconceivable to humans, "for man lives in time, in succession, while the magical animal lives in the present, in the eternity of the instant."

This extended digression affords an answer to an inevitable question, What is temporality for? The AI model contrasts offer a distinct answer, one that is less emphasized in most discussions of subjective time: Humans use temporality as a continuous compare-and-contrast operation. The circle of immediate awareness is variable and often sketchy, but it is implicitly crosschecked every second. The evolutionary advantages of immediate and continuous self-monitoring are apparent. Without this omnipresent self-awareness, even elaborate verbal productions would be guesses, however glib.

The pure experience of time

The discussion above, if accurate, consolidates a number of GPT's chronic errors, suggesting they are various manifestations of a single deficit. A similar consolidation occurs in Valtteri Arstila's "dynamic snapshots" (Arstila, 2018). Valtteri consolidates several intriguing experiments around a single model of temporality, a model situated between two others. On the one hand, the "cinematic" model denies any direct subjective experience of time. Instead, we experience a flow of sensory contents, each a snapshot, like frames in a movie, and everything

temporal is an inference from the immediately present sensorium and background knowledge. I agree that this model is much too thin – we do have temporal *experiences*, distinct from inferences about future and past. These experiences are non-sensory (Husserlian apprehensions (Husserl, 2010)). The other model, though, is more capacious. Valtteri contrasts his view with two variants of the "specious present." I agree that a "present moment" that is extended in time is ultimately incoherent. But the other variant he considers is straightforwardly Husserlian, the "retentionalist" model. The experienced present is a knife-edge, but one that includes explicit content representing the immediate past and immediately expected future. The question is, how explicit are these retentions? Valtteri reviews experiments (and illusions) affording "pure temporality," that is, a feeling of movement, succession, or change in the absence of actual perceived change.

Perhaps the simplest example of pure temporality is the experience of duration. Waiting… for the train to arrive, the light to change, the guests to appear, or just plain viewing a static scene. We're aware of the unwinding extension of the passing moments, moments however that pass without any sensory change. Such experiences are commonplace but might be misdescribed as pure duration, for the reason that our minds are always in motion (and our eyes as well). The stasis of the perceived scene may be approximately true of the physical environment, but the phenomenology is an impure mix of small changes. What we might describe as blank boredom is phenomenally dense, rather than a pure swelling void.

Valtteri focuses on other experiences of pure temporality: the experience of motion in the absence of perceived change, as in the waterfall illusion. The pure experience of causation, spontaneously felt when events are in the right temporal relationship. And the pure experience of succession, which at short time scales can be dissociated from the awareness of which of two events occurred first (Arstila, 2018).

These pure temporal experiences are distinct from Husserlian retentions in their relative lack of explicit content. The experience of a falling snowflake, according to this view, has two components. One is the change of position of the flake, while the second is the pure awareness of motion in the visual field. The dynamic snapshot is the immediate perceptual scene plus the evoked experiences of pure temporality. Valtteri asserts that the snapshot and attendant pure temporality are sufficient to account for the phenomenology of subjective time. Evaluating his theory is beyond the scope of this chapter, but the focus on pure temporality introduces a useful further articulation in theories of temporality.

In this context, then, we can return to the contrast between RNNs and GPTs with new questions. In particular, what is the computational version of pure temporality? Do either of the main neural network architectures support (or refute) the possibility of pure time?

Pure temporality is relatively content-free, but signals crucial perceptual change: Even before one knows what has changed, one knows that there has been change, and this immediately triggers an attentive search for what has altered. GPTs, as in the example above, are cheerfully blind to one prominent change, namely their own inconsistency from moment to moment. That

is, their assertions are changing in manifestly contradictory directions, but no alarm bell directs their self-examination.

Socrates famously claimed that the unexamined life was not worth living. Generative Pretrained Transformers might do well to listen to Socrates' advice, but not for its contribution to a satisfying, or ethical, life. Rather, this essay has highlighted a continuous and transient form of self-examination, basic to humans but missing in GPTs. This continual self-monitoring has a conspicuously epistemic function, offering a running check on reality. The cross-check spans brief intervals, but their overlapping sustains a consistent lived world.

Evolution would favor efficiency in this process, as in all biological processes. In the hurly-burly of experience, brains need to know promptly if there's trouble. This suggests a role for pure temporality, as the alarm bell that grabs attention to a possible problem. If, as Valtteri conjectures, pure temporality arises through dedicated, encapsulated neural mechanisms, then it could serve as the quick alert, to be followed by a more careful review involving more sophisticated cognitive capacities. A similar strategy is often cited with respect to interactions between the limbic system, especially the amygdala, and cortex (Pessoa, 2010). The amygdala seems to be dedicated to quick threat detection (Snake!), rapidly followed by higher cortical analysis (Relax, it's just a stick.). Threat detection is always an urgent matter, but so is the detection of change, especially inconsistency. Such detections warn a system that something may be wrong, and motivate a search for the cause of the discrepancy.

Toward temporality in GPTs

This discussion may suggest some advice on the future development of intelligent architectures. Recurrent neural networks involve a thorough and total retention of each moment of experience, a heavy computational task (and phenomenologically weighty as well). GPTs enlarge their cognitive horizons by eliminating recurrence, but at the price of the many failures of self-oblivion. Could there be a middle way? The existence of pure temporal experiences invites us to look for a quick-and-dirty retentional cross-check, preceding any awareness of retentional contents. This arguably requires reentrant processing, but perhaps at some primitive or early level in the cascade of activation. This is compatible with Valtteri's suggestion that pure temporality is mediated by dedicated, encapsulated, and domain specific modules. Valtteri's work clarifies the minimal information needed for subjective time in its AI analogue. LLMs could make good use of a simple signal announcing change and thereby alerting the system to a need to confirm its consistency over time. Given the manifest failures of AI in so many elementary scenarios, it could be a worthy engineering goal to install in AI systems a bit of subjective time. Such mechanisms could intervene early in the cascade of neural processing, or not, but in any case their job is simply to direct attention to change (including motion and succession). In discourse, then, such a change might be from one assertion to its contradiction, as in the GPT example above. The retentionalist model holds that contents are explicitly encoded and retained in order as time passes. This discussion, following Valtteri's lead, suggests that this elaborate and changing content is

unnecessary and accordingly maladapted. Instead, an alarm sounds when there's change; the alarm directs us to compare the present to an item in working memory. As far as subjective time is concerned, a pure temporal monitor is all you need.

Stepping back still further, we see again for humans and AI alike the radical and foundational role of the experience of time. In Valtteri Arstila's work, we find many excursions in this fascinating domain (Arstila, 2012, 2015, 2017, 2018; Arstila & Lloyd, 2014). (Here I've discussed just one of several papers and volumes on time.) There will be more, a welcome subjective future of research and discovery.

Acknowledgements

For many conversations on time and AI, I thank Cheryl Greenberg. Three paragraphs in this paper are adapted from (Lloyd, 2024).

REFERENCES

Arkoudas, K. (2023, August 9). GPT-4 Can't Reason. *Medium*.

https://medium.com/@konstantine_45825/gpt-4-cant-reason-2eab795e2523

Arstila, V. (2012). Time Slows Down during Accidents. *Frontiers in Psychology*, *3*.

https://doi.org/10.3389/fpsyg.2012.00196

Arstila, V. (2015). Keeping postdiction simple. *Consciousness and Cognition*, *38*, 205–216.

https://doi.org/10.1016/j.concog.2015.10.001

Arstila, V. (2017). Experience and the pacemaker-accumulator model. *Journal of Consciousness Studies*, *24*(3–4), 14–36.

Arstila, V. (2018). Temporal Experiences without the Specious Present. *Australasian Journal of Philosophy*, *96*(2), 287–302. https://doi.org/10.1080/00048402.2017.1337211

Arstila, V., & Lloyd, D. (Eds.). (2014). *Subjective Time: The Philosophy, Psychology, and Neuroscience of Temporality*. MIT Press.

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., & Evans, O. (2024). *The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A"* (arXiv:2309.12288). arXiv. https://doi.org/10.48550/arXiv.2309.12288

Dennett, D. C. (2023). *I've Been Thinking*. W. W. Norton & Company.

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., West, P., Bhagavatula, C., Bras, R. L., Hwang, J. D., Sanyal, S., Welleck, S., Ren, X., Ettinger, A., Harchaoui, Z., & Choi, Y. (2023). *Faith and Fate: Limits of Transformers on Compositionality*.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/10.1016/0364-0213(90)90002-E

Husserl, E. (1964). *The Phenomenology of Internal Time-Consciousness* (Fourth Printing edition). Indiana University Press.

Husserl, E. (2010). *Thing and Space: Lectures of 1907* (R. Rojcewicz, Trans.; Softcover reprint of hardcover 1st ed. 1998 edition). Springer.

Lloyd, D. (2004). *Radiant Cool: A Novel Theory of Consciousness* (1st edition). A Bradford Book.

Lloyd, D. (2024). What is it like to be a bot? The world according to GPT-4. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1292675

Marcus, G. (2020). *GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about*. MIT Technology Review. https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/

Marcus, G. (2022, March 10). Deep Learning Is Hitting a Wall. *Nautilus*. https://nautil.us/deep-learning-is-hitting-a-wall-238440/

Marcus, G. (2023, March 15). GPT-4's successes, and GPT-4's failures [Substack newsletter]. *The Road to AI We Can Trust*. https://garymarcus.substack.com/p/gpt-4s-successes-and-gpt-4s-failures

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). *Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve* (arXiv:2309.13638). arXiv. https://doi.org/10.48550/arXiv.2309.13638

Pessoa, L. (2010). Emotion and Cognition and the Amygdala: From "what is it?" to "what's to be done?" *Neuropsychologia*, *48*(12), 3416–3429. https://doi.org/10.1016/j.neuropsychologia.2010.06.038

Stechly, K., Marquez, M., & Kambhampati, S. (2023). *GPT-4 Doesn't Know It's Wrong: An Analysis of Iterative Prompting for Reasoning Problems* (arXiv:2310.12397). arXiv. https://doi.org/10.48550/arXiv.2310.12397

Tong, S., Jones, E., & Steinhardt, J. (n.d.). *Mass-Producing Failures of Multimodal Systems with Language Models*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Wittmann, M. (2016). *Felt Time: The Psychology of How We Perceive Time* (E. Butler, Trans.).

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., & Kim, Y. (2024). *Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks* (arXiv:2307.02477). arXiv. https://doi.org/10.48550/arXiv.2307.02477