Superficial Consciousness Hypothesis for Autoregressive Transformers

Yosuke Miyanishi[1,*], **Keita Mitani**[1]

[1]CyberAgent Inc.
[*]Corresponding Author
miyanishi_yosuke@cyberagent.co.jp

Abstract

The alignment between human objectives and machine learning models built on these objectives is a crucial yet challenging problem for achieving Trustworthy AI, particularly when preparing for superintelligence (SI). First, given that SI does not exist today, empirical analysis for direct evidence is difficult. Second, SI is assumed to be more intelligent than humans, capable of deceiving us into underestimating its intelligence, making output-based analysis unreliable. Lastly, what kind of unexpected property SI might have is still unclear. To address these challenges, we propose the Superficial Consciousness Hypothesis under Information Integration Theory (IIT), suggesting that SI could exhibit a complex information-theoretic state like a conscious agent while unconscious. To validate this, we use a hypothetical scenario where SI can update its parameters *at will* to achieve its own objective (*mesa*-objective) under the constraint of the human objective (base objective). We show that a practical estimate of IIT's consciousness metric is relevant to the widely used perplexity metric, and train GPT-2 with those two objectives. Our preliminary result suggests that this SI-simulating GPT-2 could simultaneously follow the two objectives, supporting the feasibility of the Superficial Consciousness Hypothesis.

Code — https://github.com/HireTheHero/PhiMesaSI

Introduction

The alignment between human objectives and machine learning models built on the objectives is a crucial yet challenging problem for achieving Trustworthy AI. Preparing for superintelligence (SI) is more challenging for several reasons. First, as we believe SI does not exist today, performing the empirical analysis for direct evidence is difficult. Second, we need to assume that SI is more intelligent than us. This implies that SI might be capable of deceiving us in conversation that they are not that intelligent; in other words, concluding by SI's output (e.g., chat log) is difficult, requiring the *intrinsic* evaluation or the evaluation of SI's internal states. Lastly, what type of unexpected property SI might have is still unclear. Here, we show our approach to these problems.

To empirically analyze the alignment between a human objective and that of SI, we make a practical assumption about the model architecture and evaluation. Specifically, we assume autoregressive Transformer (Vaswani et al. 2017), the de facto standard model in natural language processing, backbones SI, and is evaluated by standard perplexity metric (Jelinek et al. 1977). In addition, we propose *mesa-optimization* (Hubinger et al. 2021), often associated with the misalignment risk, as a key factor of our simulation. Mesa-optimization is defined as an optimization to a learner's objective (*mesa-objective*) which is different from the human objective (*base objective*). For SI analysis, we assume that SI can design the mesa-objective *at will* if it does not conflict with the base objective (otherwise it would be *corrected*). Together we assume that SI tries to set mesa-objective while keeping track of perplexity as a *base metric* (an evaluation metric for base objective).

To refrain from output analysis, we take the information-theoretic approach. In combination with the mesa-optimization framework, we assume that SI implements an information-oriented metric to update itself for its own purpose while keeping track of the original objective. This assumption allows intrinsic evaluation of the simulated SI via loss analysis, without relying on its output.

Finally, we choose consciousness as a property of interest. Although the functional role of consciousness is still unclear, several lines of work are tackling this problem (e.g., Juliani et al. (2022)). We argue that an extremely competent system (SI) could *read* the description of existing theories, and conclude that the incremental consciousness level matches the purpose. For example, when SI is facing a challenging task about episodic memory (Fountas et al. 2024) , and it recalls the theory about the relevance between consciousness and episodic memory (Budson, Richman, and Kensinger 2022), it is logically consistent for it to acquire consciousness for episodic memory. Since we assume a mesa-objective as the available tool for SI, we hypothesize that it follows an information-theoretic theory for consciousness—Information Integration Theory (IIT). Here we formally and empirically show that the consciousness metric in IIT can be used as a mesa-objective when the perplexity is set as a base metric.

Altogether, our contribution can be summarized as:

1. We propose the Superficial Consciousness Hypothesis stating that autoregressive Transformer-based SI could exhibit a complex state like a conscious agent while unconscious.
2. To the best of our knowledge, this work is the first to in-

troduce IIT analysis to the Transformer models, allowing the token-wise intrinsic evaluation.

3. We perform the pioneering mesa-optimization analysis in line with the emerging empirical quest for evidence supporting this framework (e.g., von Oswald et al. (2023).

Preliminaries

Information Integration Theory

Information integration theory (IIT; Tononi (2004)) defines the consciousness level as an information-theoretic metric Φ, given the system's cause-effect state. To summarize IIT, the complexity of a system S determines its potential consciousness level denoted as φ. To see if the system is a conscious entity, IIT cuts the system into bipartition $\mathcal{B}$ producing two subsystems $\{M_1, M_2\}$ to calculate the subsystem's φ. Finally, once the most informative (most information-reducing) bipartition or minimum information bipartition $\mathcal{B}^{\text{MIB}}$is identified, the $S - \mathcal{B}^{\text{MIB}}$ difference of φ is defined as the overall metric Φ indicating the consciousness level.
Given mutual information $I(\cdot;\cdot)$ (Shannon 1948), Mediano et al. (2022) formulated the practical estimates $\hat{\varphi}$ and $\hat{\Phi}$ of φ and Φ at time t to time τ as:

$$\begin{aligned} \hat{\varphi}_t[S;\tau,\mathcal{B}] &= I(S_{t-\tau};S_t) - \sum_{k=1}^{2} I(M^k_{t-\tau};M^k_t) \\ \hat{\Phi}_t[S;\tau] &= \hat{\varphi}_t[S;\tau,\mathcal{B}^{\text{MIB}}] \\ where\ \mathcal{B}^{\text{MIB}} &= \arg\min_{\mathcal{B}} \frac{\hat{\varphi}[S;\tau,\mathcal{B}]}{K(\mathcal{B})} \end{aligned} \tag{1}$$

where $K(\mathcal{B})$ is a penalizing term for the large bipartition. For $\hat{\varphi}_t[\cdot]$ and $\Phi_t[\cdot]$, hereafter we use the notation without the input variables ($\hat{\varphi}_t$ and $\hat{\Phi}_t$, respectively) interchangeably. We use this practical version of IIT for the rest of the paper unless stated otherwise.

Autoregressive Transformer

Core Component The core component of a Transformer model is dot-product attention $Attn(\cdot)$ with the linear weights $\{W_{MatrixType}|MatrixType \in \{Q,K,V\}\}$ with depth d followed by L linear projection layers $\{W_l|l \in \{1,...,N\}\}$. Given input X (e.g. a document to be classified), output Y^{trn} (e.g., predicted probability for a label) is calculated as:

$$\begin{aligned} Q = XW_Q, K = XW_K, V = XW_V \\ Attn(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d}})V \\ Y^{trn} = \prod_{l=1}^{L} W_l Attn(X) \end{aligned} \tag{2}$$

For simplicity, we omit the notion of *multi-heads*[1] in the attention and the bias term in the linear projection. Hereafter, we denote the computation of Eq. 2 as $Y^{trn} = Trn(X)$ for brevity. As you can see, all computations consist of a feed-forward network without recurrent connections. Some previous works provided other variants or interpretations (e.g., Oren et al. (2024)), but we leave the consideration to future work.

Algorithm Parallel to the success of ChatGPT[2] and other web applications, most state-of-the-art models solve an autoregressive task. To generate the response sequence, the model samples the next token based on the Transformer model's output and concatenates it with the existing context to move to the next iteration. Alg. 1 summarizes the algorithm. After the core component returns its output Y^{trn},

Algorithm 1: Autoregressive Transformer Algorithm.

Input: Context C
Parameter: Transformer Trn, Sampling Function $Sample$, Maximum Length T
Output: Response with Context x

$x \leftarrow C$ {Initialization}
for $t \leftarrow 1$ to T **do**
  $Y^{trn} \leftarrow Trn(x)$ {Transformer}
  $r_t = Sample(Y^{trn})$ {Sampling}
  $x \leftarrow \{x, r_t\}$ {Concatenation}
  if $r_t =$ [EOS] **then**
    $break$ {Stop with *End Of Sentence* token}
  end if
end for
return x

the generated token r_t is sampled deterministically (e.g., greedy search) or probabilistically (e.g., multinomial sampling). The sampled token r_t is concatenated to form the response x together with the context C given by the user.

Objective and Evaluation Given the context with previously concatenated responses $x_{<t} = \{C, r_1, ..., r_{t-1}\}$ at time t as an input, the training objective $\mathcal{L}$ of an autoregressive Transformer is to maximize the predicted probability of the next token r_t (Lee 2023).

$$\mathcal{L} = P(r_t|x_{<t}) \tag{3}$$

Autoregression performance is typically evaluated by perplexity PPL (Jelinek et al. 1977) [3].

$$PPL(P(R_t|X_{<t})) = \exp\left[\frac{1}{N}\log\{P(R_t|X_{<t})\}\right] \tag{4}$$

Here the large character denotes a set of its small counterparts in the dataset (e.g., X_t is a set of x_t in all the documents D), and N is the number of samples.

[1] The attention matrix in actual computation is split to the heads for enhancing the parallelism. Please refer to Vaswani et al. (2017) for more details.

[2] https://chatgpt.com/

[3] Note that perplexity is typically evaluated using another model, but we assume self-evaluation due to SI's high secrecy and competency.

Superficial Consciousness Hypothesis

Implicit Presupposition of IIT

IIT requires the inherent temporary transition of a system's internal states. Therefore, a system without recursive computation $Rec(\cdot)$, or the state update based on the previous state, is not considered conscious regardless of its complexity. Formally, to be the subject of IIT analysis, the state s_t at time t should be calculated as a function of the input x_t.

$$s_t = Rec(x_t) \tag{5}$$

Superficial Consciousness

The Transformer model does not involve recursive computation; thus, $\hat{\varphi}$ is not computable. As a system driven by an autoregression algorithm (Alg. 1), however, its state transition can be defined as:

$$s_t = Sample(Trn(x_t)) \tag{6}$$

Accumulating this state over time, $\hat{\varphi}$ is computable. Note that the *probabilistic* state transition required by the original φ might be implemented by probabilistic sampling, which we will explore in future work.

Still, we argue that $\hat{\varphi}$ computed here is *superficial* for two types of *non-intrinsicality*:

1. *Mathematical* Non-Intrinsicality: As with Alg. 1, s_t in Eq. 4 composes the input in the next time step $s_t \in x_{t+1}$. In contrast, a state of IIT's interest (e.g. a state of a human brain or a recursive system) should be *decoded* (by verbal report, locomotive action, or projection to the predicted probability) to interact with the environment.
2. *Existential* Non-Intrinsicality: Mathematical Non-Intrinsicality comes from the fact that Alg. 1 is decomposed into two main components without disrupting the other: Transformer and sampling method. Arguably, this is not the case for the human brain or recursive system—say, the motor cortex is inseparable from the rest of the brain (Sanes, Jerome and Donoghue, John 2000).

Indeed, the original IIT states that the conscious being must be subject to the criteria they call *postulates*. One of the criteria is *Intrinsicality*, defined as "its cause-effect power must be intrinsic: it must take and make a difference *within itself*" (Albantakis et al. 2023). If the autoregressive Transformer breaks this postulate (i.e., not conscious), yet its cause-effect state is mature enough to measure a certain level of $\hat{\Phi}$, we argue that SI could *behave like* a conscious agent even if it is not.

Formal Relationship between Perplexity and IIT

Since the previously sampled tokens are concatenated to form the current state, we can see that $X_{<t}$ is equivalent to a set of the states S_{t-1}. If we set the shortest time window $\tau = 1$[4], we obtain

$$\hat{\Phi}_t[X;1] = I(X_{<t}; X_t) - \sum_{k=1}^{2} I(M^k_{<t}; M^k_t) \tag{7}$$

[4]Generalizable to arbitrary τ, but $\tau = 1$ best aligns with the practical setting.

as our mesa-objective. When the system has significant $\hat{\Phi} \gg 0$, its state also has significant mutual information.

$$\begin{aligned} \hat{\Phi}_t &\gg 0 \\ I(X;Y) &\gg \sum_{k=1}^{2} I(M^k_{<t}; M^k_t) \\ \hat{\Phi}_t &\simeq I(X;Y) \\ &= H(Y) - \frac{1}{N}\log\{P(Y|X)\} \end{aligned} \tag{8}$$

$H(\cdot)$ denotes the entropy. As the second term in the last row is identical to the negative logarithmic of the perplexity (base metric; Eq. 4), maximizing $\hat{\Phi}$ (mesa-objective) could result in minimizing the base metric. In practice, we take the sum of the base metric and mesa-objective (perplexity and $\hat{\Phi}$) in the optimization and show the empirical relationship in the Experiment section.

Experiment

Experimental Settings

To validate our scenario, we train GPT-2 Medium model (Radford et al. 2019) on HuggingFace PyTorch framework with the standard WikiText corpus (Merity et al. 2017) on NVidia A40 GPU. We use batched training with 8 samples for a single epoch for the interest of training cost. MIB exploration is performed in the Optuna framework (Akiba et al. 2019), omitting the parameter $K(\mathcal{B})$ to avoid the predominant effect of a choice of this parameter.

Preliminary Result

We show that the base and mesa metrics are highly correlated in the training phase (Fig. 1), validating our mesa-optimization framework. The negative $\hat{\Phi}$ indicates that GPT-2 does not have enough cause-effect power to behave like a conscious agent, potentially due to its limited capacity.

Discussion

Here we proposed the Superficial Consciousness Hypothesis, pointing out that SI might maximize the consciousness metric while remaining unconscious in human-oriented criteria. We also showed the preliminary simulation result, marking a first step toward the information-theoretic risk analysis for SI. Although it requires an intuitive leap, we argue this is not unrealistic considering the recent trend of autonomous agents in complicated fields like academic research (Lu et al. 2024). For generalizability, our future analysis should include open-sourced model variants and more diverse datasets. We should also test our hypothesis with the original IIT framework, tackling the intractability. From a neuroscientific perspective, SI analysis could be a good testbed for the Φ metric. Uniting multiple theories, such as that of the role of consciousness on intelligence (Juliani et al. 2022), could lead to deeper insights. Cross-disciplinary collaboration should help acknowledge the significance of information-theoretic risk assessment towards post-singularity symbiosis.

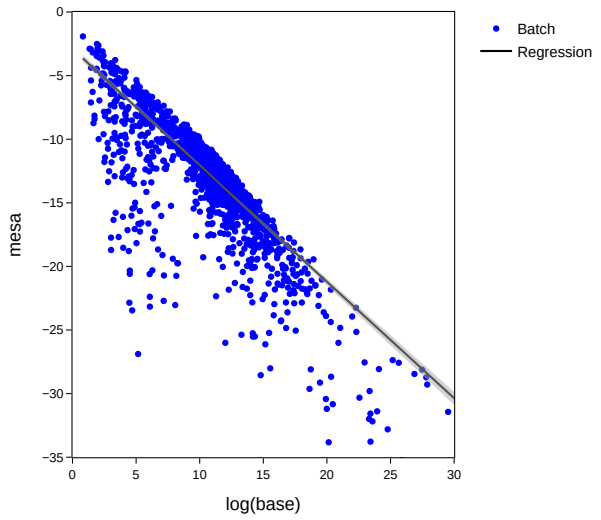

Figure 1: Correlation between the logarithmic of base and mesa metric. Each dot represents a batch. Black line with shadow indicates the ordinary least square result ($y = (-0.92 \pm 0.01)x - (2.90 \pm 0.15)$). The Granger causality test (without lag) indicates the significant predictive power of mesa metric over base metric (F value 293, $p < 0.01$).

Conclusion

Our study introduces the Superficial Consciousness Hypothesis and provides preliminary evidence through simulation for the information-theoretic risk analysis of SI. We believe this framework could offer valuable insights into consciousness and SI risk.

Acknowledgments

This work is part of the CyberAgent Seminar activity. We thank Dr. Tetsuro Morimura and Dr. T.Y. for their insightful comments. We thank ALIGN network[5] for inspiring discussion.

References

Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631. Anchorage AK USA: ACM. ISBN 978-1-4503-6201-6.

Albantakis, L.; Barbosa, L.; Findlay, G.; Grasso, M.; Haun, A. M.; Marshall, W.; Mayner, W. G. P.; Zaeemzadeh, A.; Boly, M.; Juel, B. E.; Sasai, S.; Fujii, K.; David, I.; Hendren, J.; Lang, J. P.; and Tononi, G. 2023. Integrated Information Theory (IIT) 4.0: Formulating the Properties of Phenomenal Existence in Physical Terms. *PLOS Computational Biology*, 19(10): e1011465.

Budson, A. E.; Richman, K. A.; and Kensinger, E. A. 2022. Consciousness as a Memory System. *Cognitive and Behavioral Neurology*, 35(4): 263–297.

Fountas, Z.; Benfeghoul, M. A.; Oomerjee, A.; Christopoulou, F.; Lampouras, G.; Bou-Ammar, H.; and Wang, J. 2024. Human-like Episodic Memory for Infinite Context LLMs. *arXiv preprint*.

Hubinger, E.; van Merwijk, C.; Mikulik, V.; Skalse, J.; and Garrabrant, S. 2021. Risks from Learned Optimization in Advanced Machine Learning Systems. *arXiv preprint*.

Jelinek, F.; Mercer, R. L.; Bahl, L. R.; and Baker, J. K. 1977. Perplexity—a Measure of the Difficulty of Speech Recognition Tasks. *The Journal of the Acoustical Society of America*, 62(S1): S63–S63.

Juliani, A.; Arulkumaran, K.; Sasai, S.; and Kanai, R. 2022. On the Link between Conscious Function and General Intelligence in Humans and Machines. *Transactions on Machine Learning Research*.

Lee, M. 2023. A Mathematical Interpretation of Autoregressive Generative Pre-Trained Transformer and Self-Supervised Learning. *Mathematics*, 11(11): 2451.

Lu, C.; Lu, C.; Lange, R. T.; Foerster, J.; Clune, J.; and Ha, D. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *arXiv preprint*.

Mediano, P. A. M.; Rosas, F. E.; Farah, J. C.; Shanahan, M.; Bor, D.; and Barrett, A. B. 2022. Integrated Information as a Common Signature of Dynamical and Information-Processing Complexity. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 32(1): 013115.

Merity, S.; Xiong, C.; Bradbury, J.; and Socher, R. 2017. POINTER SENTINEL MIXTURE MODELS. In *5th International Conference on Learning Representations (ICLR 2017)*. Toulon, France.

Oren, M.; Hassid, M.; Adi, Y.; and Schwartz, R. 2024. Transformers Are Multi-State RNNs. *arXiv preprint*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models Are Unsupervised Multitask Learners. *preprint*.

Sanes, Jerome; and Donoghue, John. 2000. Plasticity and Primary Motor Cortex. *Annual Review of Neuroscience*, 23(Volume 23, 2000): 393–415.

Shannon, C. E. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3): 379–423.

Tononi, G. 2004. An Information Integration Theory of Consciousness. *BMC Neuroscience*, 5(1): 42.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. In *Thirty-First Annual Conference on Neural Information Processing Systems*. Long Beach, CA, USA.

von Oswald, J.; Niklasson, E.; Schlegel, M.; Kobayashi, S.; Zucchet, N.; Scherrer, N.; Miller, N.; Sandler, M.; y Arcas, B. A.; Vladymyrov, M.; Pascanu, R.; and Sacramento, J. 2023. Uncovering Mesa-Optimization Algorithms in Transformers. *arXiv preprint*.

[5]https://www.aialign.net/