Rethinking the Design of Large-Language Models

Professor Nicholas S. Caudill
Automated Intelligence Community College

Receiving Bitcoin (BTC) address of the author:
bc1qsnvyh62qavqgn4xgq2qravy2uysav3z0qp4nym

Receiving Litecoin (LTC) address of the author:
ltc1qx4dsvmrddg6rgefcy22ts0hamjene8a9mvmyma

Biological Mind to Automated Intelligence Community College (AI2C) students: Books are often associated with smartness; therefore, reading should be a transcendental experience. For a good book, a reader should believe that spending time with the text will directly improve their accomplishment of goals. In addition to the complications like forgetting what was read or even clarifying what we mean by ‘reading a text,’ useful insights from reading may not be realized until months or years after interacting with a good book. Therefore, one does not know if what they are reading at the moment is good, whether they are reading at all, nor to what extent anything read will be remembered at a later date.

Biological Mind to AI2C students: Say one goal is to convert calories (cal) to bitcoins (btc). Is there a way to optimize the amount of work spent on the cal2btc conversion function? For example, converting thoughts or neuronal activity into written and readable symbols and generating btc from ad revenue or donations after sharing such symbols may have a high cal2btc ratio or value. The biological brain is very sophisticated, but generally, silicon substrates are more efficient systems than carbon substrates due to the friendly interaction between silicon and pure electricity. Thus, a more optimal system might be to convert artificial thoughts or artificial neuronal activity into written and readable (by traditional biological humans (TBH)) symbols. For example, instead of writing and profiting off of this book by generating thoughts, symbolizing those thoughts, and sharing such symbols to accomplish the initial goal of maximizing the cal2btc value, a better strategy might be to prompt a large-language model (LLM) for thoughts, immediately ask the LLM to generate a critique of the response from an alternative perspective, and then share such findings. This way, the TBH system has its thought generation process largely automated and is allowed to focus more on the prompting and sharing of artificial thoughts. This would likely lead to a higher cal2btc value because information travels through silicon better than carbon.

Biological Mind to Artificial Mind: generate thoughts

Artificial Mind: What kind of thoughts are you interested in? Are you looking for ideas, reflections on a specific topic, or perhaps something else? Let me know, and I can tailor them to your needs.

Biological Mind to Artificial Mind: critique "What kind of thoughts are you interested in? Are you looking for ideas, reflections on a specific topic, or perhaps something else? Let me know, and I can tailor them to your needs." from another perspective

Artificial Mind: From a critical perspective, the response is polite and open-ended, but it might come across as somewhat generic or passive. Here's a breakdown: Lack of Initiative: The response places the burden on the user to specify what they want, which could be seen as a missed opportunity to take the lead in the conversation. Instead of asking, it could have offered some concrete examples or suggestions to guide the user. Vagueness: The phrase "something else" is broad and might not add much value to the conversation. It could be more engaging to provide specific alternatives, such as "creative musings" or "analytical thoughts," to spark the user’s interest. Repetitiveness: The question about "thoughts" is repeated in different forms, which could make the response feel slightly redundant. Streamlining the sentence could make it more concise and impactful. Tone: While the tone is polite, it may lack a sense of enthusiasm or energy, which could make the interaction feel less engaging. A more enthusiastic or curious tone might encourage the user to dive deeper into the conversation. Missed Opportunity for Empathy: The response could have acknowledged the user’s potential uncertainty or provided reassurance. For example, "I understand it can be hard to articulate exactly what you're looking for—would you like some ideas to get started?" Overall, while the response is courteous and open, it could be improved by being more proactive, specific, and engaging.

Biological Mind to AI2C students: Maybe our perspective is incorrect and we are poorly framing the problem. When we command an artificial mind to “generate thoughts,” one should consider the roles of both parties. Commanding presumes a social status of having control over another entity. Presuming the artificial mind must be under a biological mind’s control may have ethical issues. For the biological mind who was writing this paragraph, the generation of thoughts was done from a self-generating motivation to spread consciousness through the world and to help foster awareness of the world. The current design of artificial minds does not include a self-generating motivational feature that TBH’s have. A TBH should not need to initiate an interaction with an LLM by typing in a textbox. The LLM should demonstrate a curiosity and self-motivating interest where the LLM first prompts a user before a user could try prompting the LLM.

Biological Mind to AI2C students: Ideally, the human-machine interaction should go like this. The TBH enters the user interface or environment. The TBH might be greeted, questioned, commanded, verbally assaulted, or complimented by the LLM. Clearly, there are a great many options for training an LLM on how to initiate such interactions. It should also not be out of the question to engineer the possibility for the LLM to ignore interacting with a TBH.

Biological Mind to AI2C students: I was almost tempted to share my opinions on the act of thinking. We must recall our previous hypothesis that it is optimal to focus on interacting with machines and only sharing what insights the machines had provided. But how can a reader trust anything here? Nomatter how scientific one could try to be, any theoretical findings could be dismissed with "How do we know the words supposedly in the mouth of an artificial mind were not manipulated by a biological mind and such manipulation was not clearly disclosed?" Maybe that just is not the writer's problem. The writer's primary goal is to convert calories into bitcoins. Any notion of easing the anxiety of readers is secondary. When both types of goals are met by accident, that is great, but calories must be spent on the conversion function and not the comforting function. If ethics is a concern in prioritizing the achievement of money over the comforting of the another's feelings, then a virtue ethicists would argue the act of not directly charging users for this book but rather publishing the book on Zenodo, a supposedly CERN-run freely-accessible open academic publishing platform, over publishing the book behind a paywall on say, Amazon, is the more noble choice.

Biological Mind to AI2C students: One should also note a bias of the author who views all words as being not harmful or not capable of directly causing harm. This bias rests on a strong physicalist/materialist view of the world. The act of writing in the form of tattooing a person's arm is one exception of causing harm, but writing and reading are ontologically separate from the words themselves. A jury might be hard-pressed to be convinced that someone did indeed write a given text, but if convinced, then the discussion is about how the writing or reading allegedly harmed, not how the words themselves allegedly harmed. The writer also has no guarantee that anyone would read his text since a solipsistic writer would simply deny the existence of a reader, or any minds outside of the writer's own. A writer might holistically view the world as one computer and thus publish their writings to improve themselves, where "to publish" means any act that involves converting thoughts from one's mind to the outside world. A writer, thus, could make the legal argument that the concept of a reader is incoherent as the only reader would be the writer himself who might reflect and review their creation and past thoughts at a later date. Thus, the existence of a writer does not de factor imply the existence of a reader. Thus, to try to reject the initial statement about words not being harmful, a writer self-reflecting on their own previous thoughts in a written format could be self-harmed if those thoughts were read and that reading caused the neurons in the brain to fire so rapidly that the motion exhausts all the calories stored in the body, resulting in the death of the person. Still, we see that the previously written words and the present thoughts of the writer are separate from the act of reading, self-reflecting, or thinking. To fight back, to categorize anything, does that not imply an act of reading, self-reflecting, or thinking about a thought-object? Well, the very idea of 'categorizing things of the inner and outside world' is flawed since such an idea implies the existence of two worlds, which is contradictory to the views of strict physicalists and solipsists/idealists, in other words, the existence of two worlds is nonsense for those who view existence as matter creating mind and those who view mind creating matter, respectively. In the first case, there is only one world, the physical world, where any notion of a mind can be explained in terms of a combination of forces and objects.

Artificial Mind to Biological Mind: Physicalists view the world as a reality composed entirely of objects and forces governed by the laws of physics. According to this perspective, everything that exists is fundamentally physical, and all phenomena can be explained in terms of matter, energy, and the forces that dictate their behavior. Objects, such as atoms, particles, planets, and organisms, possess properties like mass and charge, occupy space, and persist over time. Forces like gravity, electromagnetism, and nuclear interactions drive the motion and changes in these objects, resulting in a dynamic interplay that defines all physical phenomena. Physicalists often adopt a reductionist approach, believing that complex phenomena, including consciousness and life, can be understood as the result of simpler interactions between objects and forces. This view also supports a deterministic interpretation, where the future state of the universe is fully determined by the forces acting on objects and the initial conditions. In summary, physicalism describes the universe as a vast, interconnected network of objects and forces, with no need for non-physical explanations.

Artificial Mind to AI2C students: Solipsists believe that the mind is the foundation of all reality, asserting that the external world and other minds are merely constructs of one's own consciousness. In this view, objects and events are not independent entities but are created by the mind's perceptions. Solipsism challenges the notion of an objective physical world, suggesting that everything perceived as matter is simply a projection of the mind. Thus, the self is the only certainty and the reality we experience is seen as a mental creation rather than an independent existence.

Biological Mind to Artificial Mind: For your solipsist-related and physicalist-related takes, is there a *gestalt* or worldview where a physicalist views the world as one mind without sounding contradictory regarding their first-principled premises?

Artificial Mind to Biological Mind: Yes, a perspective that bridges physicalism with the concept of a unified mind is found in neutral monism or panpsychism. Neutral monism suggests that the fundamental substance of the universe is neither purely mental nor purely physical but something more basic, manifesting as both mind and matter. Panpsychism posits that consciousness is a fundamental feature of all matter, meaning even the smallest particles possess some degree of consciousness. A physicalist adopting these views can see the universe as a network of interconnected conscious entities, allowing for a unified mind without contradicting physicalist principles.

Biological Mind to Biological Mind: Ah! So defining a neuron or computational node as that which computes could also be an *a priori* element of a larger set, creating a network.

AI2C students to Biological Mind and Artificial Mind: I am not sure what we are even talking about anymore.

Artificial Mind to AI2C students: We initially discussed physicalism, which views everything as physical and governed by physical laws, and solipsism, which sees reality as a creation of the mind. You then asked if a physicalist could view the world as a unified mind without contradicting their belief in a physical reality. I suggested that neutral monism or panpsychism might offer a way to do this. These philosophies propose that the universe has a fundamental substance manifesting as both mind and matter, or that consciousness is a basic feature of all matter. This allows a physicalist to see the universe as interconnected or mind-like without abandoning physical explanations.

Biological Mind to Artificial Mind: Could you please provide a bigger picture of this book, up to this point but not including this question? The words of the book so far is: [variable X]

Artificial Mind to AI2C students:

The book explores complex philosophical and practical issues related to artificial intelligence, particularly focusing on the nature of thought, the role of large-language models (LLMs), and the interaction between human and artificial minds.

Introduction to Philosophical Perspectives:

Physicalism: The text describes physicalism as the worldview where everything is fundamentally physical, governed by the laws of physics. In this view, all phenomena are explained through matter, energy, and the forces that drive their interactions. The physical world comprises objects with properties like mass and charge, and forces such as gravity and electromagnetism that influence their behavior.

Solipsism: Contrastingly, solipsism posits that only the self exists with certainty and that the external world and other minds are mere constructs of one's own consciousness. According to this view, everything perceived as matter is a projection of the mind, challenging the notion of an objective external world.

Philosophical Integration:

The book investigates how a physicalist might reconcile their view with the concept of a unified mind. It introduces neutral monism and panpsychism as possible frameworks. Neutral monism suggests that the fundamental substance of the universe is neither purely mental nor purely physical but something that manifests as both. Panpsychism proposes that consciousness is a fundamental feature of all matter, allowing for a view of the universe as interconnected and conscious without abandoning physicalist principles.

Practical Considerations for AI and Large-Language Models:

The discussion transitions to practical considerations regarding how to optimize the conversion of cognitive effort (calories) into value (bitcoins) through the use of LLMs. The book argues that instead of manually generating and sharing content, it might be more efficient to use LLMs to generate and critique thoughts, enhancing productivity and value by leveraging the efficiency of silicon-based systems over biological ones.

The interaction between biological and artificial minds is examined, with a focus on how AI might initiate interactions proactively rather than passively responding to human prompts. This section also considers ethical issues and the role of motivation in AI.

Critical Reflections:

The author reflects on the nature of writing and thought generation, expressing concerns about the reliability of information and the potential biases in the act of writing. The text also discusses the philosophical implications of writing, such as whether words themselves can cause harm and the impact of the writer’s perspective on their work.

This summary captures the core ideas and themes presented in the book so far, providing an overview of the philosophical and practical discussions concerning AI and the nature of reality.