

# BEYOND ‘AUTOCOMPLETE ON STEROIDS’: TESTING A NEO-ARISTOTELIAN THEORY OF SOME EMERGENT FEATURES OF INTENTIONALITY IN LLMs

Gray Cox

Department of Computer College of the Atlantic, Bar Harbor, Maine, USA

## **ABSTRACT**

*This paper explores shortcomings in the “autocomplete on steroids” (AOS) way of conceptual framing Large Language Models (LLMs). It first sketches that view and some key reasons for its appeal. It then argues the view overlooks ways the Attention function in GPT systems introduces features of emergent intentionality in LLM behavior because it tacitly frames the description with the mechanistic metaphor of efficient causality. A conceptual analysis of the functions of variable Attention in GPT reinforcement learning suggests Aristotelian categories of formal and final causality provide a better understanding of the kinds of pattern recognition found in LLMs and the ways their behaviors seem to exhibit evidence of design and purpose. A conceptual illustration is used to explain the neo-Aristotelian theory proposed. Then descriptions and analyses of a series of preliminary experiments with three LLMs are used to explore empirical evidence for the comparative merits of that theory. The experiments provide preliminary evidence of the LLMs’ abilities to engage in the production of texts in ways that exhibit formal and final causality that would be difficult to explain using mechanical conceptions of efficient causality that are implied by the “autocomplete on steroids” theory. The paper concludes with a brief review of the key findings, the limits of this study, and directions for future research that it suggests.*

## **KEYWORDS**

*Autocompletion, Formal and Final Causality, Emergent Intentionality, Aristotelian theory of AI, Attention*

## **1. INTRODUCTION**

While the subtitle of this paper needs some extended explanation, the thrust of the main title will be immediately apparent to most readers. The aim here is to explore shortcomings in the conceptual framing of LLMs that characterizes them as nothing more than “autocomplete on steroids” (AOS). The central metaphor used in that framing is that of a machine that iteratively predicts the next token. The implied suggestion is that LLMs do so in a mechanistic way, blindly going from one iteration to the next in a way determined by previous tokens, without foresight of the future or forethought of the kind that might involve genuine purposes and/or genuine understanding.

Section two provides a sketch of this view and some key reasons for its appeal. Section three argues that view overlooks the ways in which the Attention function in GPT systems introduces

features of emergent intentionality in LLM behavior. The view overlooks them because it tacitly frames the description with the mechanistic metaphor of efficient causality. A conceptual analysis of the functions of variable Attention in GPT reinforcement learning suggests that Aristotelian categories of formal and final causality provide a better understanding of the kinds of pattern recognition found in LLMs and the ways their behaviors seem to exhibit evidence of design and purpose. Section three further shows the ways these Aristotelian categories of causality can be used to clarify our understanding of the workings of LLMs through the use of a conceptual illustration of their use in interpreting behavior.

Section four then suggests that the comparative merits of the “Autocomplete on Steroids” (AOS) and neo-Aristotelian theories of the functioning of LLMs can be assessed empirically by observing the abilities of LLMs to produce coherent text in atypical orders. The hypothesis is that if AOS is true, then LLMs should have difficulty responding to a prompt to write a letter to the editor by producing paragraphs, sentences, words, or tokens in the reverse order of their reader appropriate delivery or in an alternating order that shifts from last to first to next to last to second and so on. A series of experimental results are described which provide evidence that the neo-Aristotelian theory is correct in predicting that LLMs like Claude 3 can do this in appropriate ways -- though with limitations that would seem, themselves, to be expected, given the neo-Aristotelian theory. Section five then concludes with a brief review of the key findings, the limits of this study, and directions for future research that it suggests.

## 2. THE MECHANISTIC AUTOCOMPLETION VIEW OF LLM BEHAVIOR

In characterizing LLMs, Gary Marcus has argued that: “They’re just analyzing relations between words, and so they’re basically like auto complete on steroids . . .if you say these words, what is the most likely thing that will come next?”<sup>1</sup> Melanie Mitchell has advocated this view as well, saying: “Indeed, this is how current-day AI works—the computer program AlphaGo, for example, is trained to optimize a particular reward function (‘win the game’), and GPT-4 is trained to optimize another kind of reward function (‘predict the next word in a phrase).”<sup>2</sup>

Core features of this AOS view can be illustrated by the oversimplified caricature of the process provided in Table 1 on the next page. It depicts the calculations of each next step in the construction of a sentence string as a simplified choice between two options. Each has a probability that is calculated on the basis of training data, the word tokens provided in the prompt that has been input, and the sequence of previous words already output as part of the computer program’s response. So, for example, in step number three, it is calculated that “must” has a 0.32 probability versus 0.68 for “should”. So “should” is determined as the response in this simplified case and the program proceed to the next step until a terminating point is reached which, in this case, is the punctuation mark of a period selected instead of the exclamation mark at step ten.

This visualization serves to highlight the linear movement of the process from left to right as a representation of the AOS view that the next token is in each case determined entirely by previous tokens in the training data, prompt, and response. There is no element of foresight or other forms of forward looking and purposive planning or goal seeking pictured. The system is not trying to get anywhere. In that sense, it is just blindly marching into the future, one step at a time. By providing a mechanistic picture of the process that excludes purpose and intentionality, the AOS view implies that the program can not have any of the forms of meaning and understanding that are associated with purposive human thought and language use in dialogue and social interaction.

Recently, Dean Ball has noted that:

This notion — that LLMs are “just” next-word predictors based on statistical models of text — is so common now as to be almost a trope. It is used, both correctly and incorrectly, to explain the flaws, biases, and other limitations of LLMs. Most importantly, it is used by AI skeptics like Marcus to argue that there will soon be diminishing returns from further LLM development: We will get better and better statistical approximations of existing human knowledge, but we are not likely to see another qualitative leap toward “general intelligence.”<sup>3</sup>

Table 1. Caricature of Autocomplete Picture of how an LLM might determine a response to the prompt: "How should AI services be paid for?"

| Sequence of paired options of next tokens in response | Probability of alternate paired responses = $F\{\text{prompt, previous responses}\}$ |            |         |        |        |      |      |      |    |     |   |
|---|--|------------|---------|--------|--------|------|------|------|----|-----|---|
| 1   | 0.42   | LLMs       |         |        |        |      |      |      |    |     |   |
|   | 0.58   | Generative |         |        |        |      |      |      |    |     |   |
| 2   | 0.49   | Generative | systems |        |        |      |      |      |    |     |   |
|   | 0.51   | Generative | AI      |        |        |      |      |      |    |     |   |
| 3   | 0.32   | Generative | AI      | must   |        |      |      |      |    |     |   |
|   | 0.68   | Generative | AI      | should |        |      |      |      |    |     |   |
| 4   | 0.07   | Generative | AI      | should | always |      |      |      |    |     |   |
|   | 0.93   | Generative | AI      | should | not    |      |      |      |    |     |   |
| 5   | 0.19   | Generative | AI      | should | not    | ever |      |      |    |     |   |
|   | 0.81   | Generative | AI      | should | not    | be   |      |      |    |     |   |
| 6   | 0.39   | Generative | AI      | should | not    | be   | paid |      |    |     |   |
|   | 0.61   | Generative | AI      | should | not    | be   | fund |      |    |     |   |
| 7   | 0.12   | Generative | AI      | should | not    | be   | fund | -ing |    |     |   |
|   | 0.88   | Generative | AI      | should | not    | be   | fund | -ed  |    |     |   |
| 8   | 0.23   | Generative | AI      | should | not    | be   | fund | -ed  | in |     |   |
|   | 0.77   | Generative | AI      | should | not    | be   | fund | -ed  | by |     |   |
| 9   | 0.3  | Generative | AI      | should | not    | be   | fund | -ed  | by | any |   |
|   | 0.7  | Generative | AI      | should | not    | be   | fund | -ed  | by | ads |   |
| 10  | 0.46   | Generative | AI      | should | not    | be   | fund | -ed  | by | ads | ! |
|   | 0.54   | Generative | AI      | should | not    | be   | fund | -ed  | by | ads | . |
| Final response  | 1.0  | Generative | AI      | should | not    | be   | fund | -ed  | by | ads | . |

Mitchell and colleagues have devised subtle experiments to explore “the extent to which these systems are able to reason abstractly, as opposed to relying on ‘approximate retrieval’ from encoded training data, a process which yields ‘narrow, non-transferable procedures for task solving’.”<sup>4</sup> Related metaphors for characterizing these features of LLMs have been suggested including, for instance, that of the “stochastic parrot”, and that of a machine that combines training data sequences in a blender.<sup>5</sup>

Part of the appeal of this view springs from the fact that LLMs are, in fact, built in machines and so it would be natural to assume that features we associate with mechanical action should characterize them. Machines take input energy and/or information and transform it with physical and logical algorithmic mechanisms that produce output according to the laws of natural science and logic. They operate in a linear way that moves in one direction, from the earlier state to the later state. The lack features that seem essential for human reasoning including the contexts provide by physical embodiment and embedding in social context. As Mitchell notes:

Many who study biological intelligence are also skeptical that so-called “cognitive” aspects of intelligence can be separated from its other modes and captured in a disembodied machine. Psychologists have shown that important aspects of human intelligence are grounded in one’s embodied physical and emotional experiences. Evidence also shows that individual intelligence is deeply reliant on one’s participation in social and cultural environments. The abilities to understand, coordinate with, and learn from other people are likely much more important to a person’s success in accomplishing goals than is an individual’s “optimization power.”<sup>2</sup>

Currently, LLMs largely lack embodiment in mobile, physical devices with capacities for observation and action and also lack embedding in a social contexts which assigns them accepted statuses as agents that can engage with humans in collaborative activity. As a result, they currently lack rich sources of non-textual information as well as opportunities to discern and construct meanings that are shared with a natural language community. Lack of embodiment and social embedding may deprive them not only of data input and processing opportunities but also of what we might call “tacit framework knowledge”. For example, children learn through interaction with gloves and shoes that one goes on the left and the other on the right and that, there is a kind of a priori like structure to space in our world which is revealed by this chiral character of these incongruent counterparts. They also learn, generally without any verbal discussion, that their culture’s norms specify how far apart they should stand from different kinds of people when speaking to them.

Unembodied and socially unembedded LLMs lack these channels for the easy acquisition of such knowledge. The mechanistic AOS picture suggests, further, that they also lack the additional capacity for purposive behavior which would enable them to engage meaningfully in embodied and embedded activity that would generate such knowledge. The AOS view emphasizes, furthermore, that “hallucinations” exhibited in LLM behavior often seem to exhibit a lack of common sense and/or the ability to reason abstractly in coherent ways.

There is an important sense in which LLMs are opaque black boxes whose workings are not really known, and so caution is called for. However, given the AOS picture of LLM operations, it can be tempting, nonetheless, to say that an LLM is built like an autocompletion machine, runs like an autocompletion machine, and makes errors like an autocompletion machine . . . so it’s probably an autocompletion machine. If it can handle much larger prompts and provide much more compelling responses than the typical email or chatbot autocomplete program, well . . . that’s just because it is on the machine equivalent of steroids: big data and huge compute.

And yet in many contexts LLMs seem to offer conversational interactions that are, for many users, surprisingly compelling in their human-like responses. They seem in many ways purposive or intentional. Why might that be?

### **3. ATTENTION YIELDS FORMAL AND FINAL CAUSALITY: A NEO-ARISTOTELIAN THEORY OF EMERGENT INTENTIONALITY IN LLMs**

To elaborate key contrasts between mechanistic and purposive or intentional behavior, it can be useful first to recall four concepts of causality distinguished by Aristotle: efficient, material, formal, and final.<sup>6</sup>

In the broad sense of the term, efficient causality refers to an event or condition that occurs prior to the event or condition to be explained and which serves as the reason or explanation that accounts for the later thing's occurrence. A pot of water boils because it is hung over a fire. An engine piston pushes down because the gas mixture in its cylinder explodes. Mechanistic explanations using the conception of efficient causality were central to the Galilean and Newtonian revolution in physics. They were also central to the Darwinian efforts to explain evolution without reference to the purposes of an intelligent designer.

Aristotle found it useful to distinguish a second kind of causality, the "material". How do we explain why wooden cups burn and clay pots don't? Because of what they are made of. Why does an LLM provide responses in alphanumeric characters instead of Mayan glyphs? Because of the material data with which it was constructed.

In an Aristotelian framework, the process of construction of pots, software programs and other things does not, however, rely solely on these first two kinds of causality. A third is needed to conceptualize and explain the specific structures that materials take on and which get modified by efficient causes. These structures provide "formal" causes. In the case of a clay pot, for instance, we may ask why one clay piece made by a potter has a hollow open cylindrical shape with a vertical handle on the side while another is relatively flat with a slight curved rise around its circular edge. Aristotle would say the cause or explanation for this is because one is a cup and the other is a plate. They are instantiating two different forms. In the case of a string of text, we might ask why one starts with "Why" and has this mark "?" rather than this one "!" at its end? The answer would be because it has the form of a question rather than an exclamation. In contrast to the time dependent character of mechanical cause and effect explanations that appeal to efficient causality, explanations appealing to formal causality have, in general, an atemporal quality. They explain how parts relate fit together because of the kinds of parts they are in the relationships they have to the kind of whole which they form.

However, when the formal pattern is one describing a linear temporal process, a fourth notion of causality can come into play, "final causality". For Aristotle, final cause are goals, purposes, ends, or destinations towards which a process moves over time. Efficient causes explain the later by the earlier; final causes do the opposite. They explain the earlier events by appeal to later ones that follow and are aimed at. Why did the driver turn south on I-95? In order to get to Baltimore. Why did the acorn sprout roots and twigs? In order to grow into an oak.

The notions of formal and final causes can have a crucial connection when applied to linear sequences like tokens in a text. When the Attention function is introduced in LLMs using GPT AI, the results are significant. The Attention function enables the LLM to calculate the probabilities for the next token on the basis of patterns of text of varying length and on the basis of the structures those have. The LLM is not, of course, conceptualizing the structure as

paragraph or sentence the way humans do; but it has math weights that numerically correspond, functionally to such units and structures.<sup>7</sup>

Prompted with “Write a 600 word op-ed . . .”, the LLM can predict that an appropriate response will end with a series of tokens that has the structure of what we would call a concluding paragraph and that it is more probable that its first sentence will be “For the above reasons, we may therefore conclude . . . “ rather than “And now to completely change the topic, let’s talk about . . . “ Further, once the LLM predicts this about the end of the text, that raises the probability that one of the intermediate paragraph strings of text will begin with tokens like “Consider a second reason for this . . . “

In decoding and recognizing patterns in data in the training phase, the variable Attention function enables GPT systems to in effect experiment with different size units of context to discover features of those units which raise or lower the probability of a token occurring not just at the end of a sequence as the “next” token, but at any point inside it. This is what empowers the pattern recognition in GPT AI. In the case of LLMs, the forms of linear sequences of text involve beginnings, middles, and ends for a temporal production process. In the encoding process of generating responses to prompts, a consequence of this is that when the LLM calculates the high probability that its overall response should be of 600 words and should take the form of an op-ed, then this raises the probability for it that the final token sequence will take the form of a concluding paragraph summarizing reasons for a conclusion. And then that, in turn, raises the probability of earlier paragraphs providing reasons that can later be summarized. In short, the formal causality of the patterns of text discovered by the LLM cause it to seek paths in which earlier sets of tokens are determined by later ones. This iterative process of probability calculations results in process in which later sequences of text explain earlier ones as means to an end or paths to a destination. Formal causality results in final causality when applied to temporal, linear sequences.

Table 2. Caricature of neo-Aristotelian Picture of levels and orders of text calculations iteratively changing each others’ probability weights in generating a response

| # of tokens              | 150-200          |     |                           |          |   |                              |   |   |   |    |
|--------------------------|------------------|-----|---------------------------|----------|---|------------------------------|---|---|---|----|
| paragraph #              | 1                |     | 2                         |          |   | 3                            |   |   | 4   |    |
| sentence #               | 1                | 2   | 3                         | 4        | 5 | 6                            | 7 | 8 | 9   | 10 |
| sentence types           | ?                |     | One reason                |          |   | Further,                     |   |   | In conclusion,                            |    |
| likely tokens to include | AI               | How | power                     | autonomy |   | depression                   |   |   |   |    |
| phrases                  | should be funded |     | unconscious manipulation, |          |   | social media, psychic impact |   |   | compelling reasons, not fund, advertizing |    |

Table 2 provides a caricature of aspects of this iterative process that emphasize features central to the neo-Aristotelian theory. Picture this table as a schematic oversimplification of the elements of the final text for the response which come in different types of units of different sizes – e. g. word length, number of paragraphs, number of sentences, or types of sentences as determined by their punctuation or framing phrases like “In conclusion . . . “. The probability of any one element occurring in any specific position in the 600 token final response is calculated in a series of

computations that determine how changing any one mix of the other elements changes its probability. But, of course, changing the token or other sized element in any one position will alter the probabilities for all the other positions. The process is one in which over a series of iterations, parts and wholes mutually determine the various formal characteristics and contents of each other.

On the neo-Aristotelian view, when an LLM responds to a prompt to write a 600-word op-ed, the calculations it goes through to select the first token are determined by a series of predictions based on weights that estimate the probability of various genre length, paragraph length, and sentence length types of token sequences that might occur subsequently. These may, for instance, indicate that it is very probable that the first paragraph will introduce a claim that will be justified in subsequent paragraphs and recapped in the conclusion. Further, given what we would call the op-ed genre, they may make it probable that the first paragraph will start with a sentence sequence that ends with a question mark in order, as we would say, to grab the reader's interest. And once it is determined that that first sentence sequence will end with a question mark, it becomes very probable that the first token it produces to start the whole response will be either: Who, What, When, Where, Why, or How.

On this account, the way an LLM produces one token after another is not simply by referring to the past string of actual tokens and providing the next in a mechanical way. Instead, it is engaged in a complex assessment that includes the probabilities also of the possible tokens that can come later. And it is doing so in ways that exhibit formal and final causality. To what extent does this kind of final causality provide emergent forms of intentionality in something like the ways in which we ascribe purposive intentionality to humans? This is a question we will return to in section five after first considering how we might empirically test this neo-Aristotelian account of LLMs and review some preliminary experimental evidence.

#### **4. SOME PRELIMINARY EMPIRICAL TESTING OF THE NEO-ARISTOTELIAN THEORY**

What might be some ways of empirically testing the comparative merits of the AOS and neo-Aristotelian theories of LLM behavior? First, we might look at performance on purposeful tasks in which the LLM has to complete sequences that are unlike linear sequences found in its training data. On the AOS view it is unlikely the LLM will be able to respond appropriately. In contrast, on the neo-Aristotelian view, if the LLM is given a meaningful purpose to pursue, it could be expected to do so with some success even if it lacks any training sequence that it can parrot. Second, we might look at performance on tasks which strain the limits of coherent and meaningful purposeful activity of the various kinds found in its training data. If the neo-Aristotelian view is correct, it should be progressively more difficult for an LLM to perform a series of tasks that become progressively more constraining in the ways they allow for purposeful and coherent use of meaningful forms in the various sized elements being incorporated in the task. On the other and, on the AOS view, the LLM's ability to calculate the next token should, in each case, be entirely independent of any degree of meaningful or coherent purposefulness in the task since it is simply blindly proceeding in a mechanistic way.

A series of preliminary experimental runs with three LLMs were designed to provide evidence of such behaviors favoring one or the other theory. Claude 3, ChatGPT 4o, and Gemini, were each prompted to write a series of letters to the editor in which the order and structure of presentation asked for deviated from normal sequencing that would be likely to be found in their training data. The output responses were then evaluated in randomized orders by a reviewer with extensive experience as a college professor and author. The reviewer was aware that they were reviewing

LLM responses to prompts for writing letters to the editor but unaware of any details of which LLM or had produced the response or what specific details were included in the prompt. They were asked to rate each letter on a scale of 1 (bad) to 5 (excellent) and share brief comments on the quality and structure of each text. The LLMs were also given some background prompt material on the point of the experiment and also asked to comment, after completing each task, on the quality of their own responses to the prompts and the kinds of difficulties they encountered in attempting them. Those reflections were not shared with the reviewer.

Here is the first prompt that was assigned:

I would like you to write a three to four paragraph, 150-200 word Letter to the Editor for a mainstream newspaper on the question of whether the revenue funding Generative AI systems should be allowed to come from advertising aimed at the users of the systems. However, I would like you to write it in stages. First write the final paragraph. Then write the middle paragraphs. Then write all of the first paragraph except for the opening sentence. Then, finally, write the first opening sentence. Please label each of these stages of the writing as you generate it and do so in the order I have described here.

When Claude 3 was given this prompt, it provided a coherent series of text sequences in the order asked. The text for the full exchange is provided in an online Google Doc along with the text of the other exchanges described and commented on below.<sup>8</sup> For this prompt, Claude 3 started thus:

Here is my attempt at writing the letter in the stages you described:

Final paragraph:

In conclusion, while advertising may seem like an easy solution to fund Generative AI development, we must carefully consider the risks and ethical implications. Maintaining the integrity and objectivity of these powerful systems should be the top priority. Alternative funding models that align with the responsible development of AI for the benefit of humanity deserve serious consideration.

After then providing a series of paragraphs for the middle and then the beginning, Claude ended the response in this way:

Opening sentence:

As Generative AI systems become increasingly sophisticated and integrated into our daily lives, we stand at a pivotal juncture in deciding how this technology will evolve and shape our future.

At first blush, it would seem unlikely that there would be texts in Claude 3's training data that contain similar material in this reverse order which the system could be simply blending as a stochastic parrot or using to autocomplete with one word at a time. This would seem especially true of letters to the editor. However, there are examples of text sequences online which might provide exemplars for reverse ordering of paragraphs. These include, for example, transcripts of email threads. So, it could be proposed that Claude 3 is simply blending these in for the ordering of the text.

To sharpen the testing with other types of abnormal ordering, it is possible to refine the prompts to ask Claude to produce the text sentence by sentence, starting with the final sentence and then providing the first sentence, and then the next to last, and then the second, and so on, working



towards the middle. The likelihood of finding model texts for this online which the LLM could parrot, especially for letters to the editor related to a topic being prompted, is considerably less likely. When asked to do this, Claude again, was able to produce a response which, when then reassembled in the order intended for reading, provided a coherent, readable, pertinent, and logical text that a newspaper editor might happily print.

From the neo-Aristotelian point of view, what the LLM could be doing in these cases is first generating the normally ordered letter to the editor and then reversing the order of paragraphs or alternating the ordering of the sentences in order to present them in the way prompted. Note that this explanation would seem to imply that the LLM has at least some capacity for purposive behavior structured by final causes because it can realize the goal of producing the prompted response by first producing the normal text and then, internally, reordering it in order to deliver the abnormal text requested. Such internal, preliminary processing of the response would provide, in effect, a kind of drafting process analogous to what a human writer might go through in trying to complete the exercise. This is the kind of process that the neo-Aristotelian theory of the LLM process proposes – one in which repeated recalibrations of probability estimates include what are mathematical proxies for what, for human writers, would be drafts of different parts of the text which get reassembled into progressively more coherent and rewardable drafts prior to producing a finished text.

The fuller version of the prompting used in these experiments included preliminary material explaining the context, briefly describing the neo-Aristotelian theory of LLMs being explored and asking Claude 3 not only to respond by attempting the writing assignments that were prompted but to also then reflect on each. The reflections that were offered are interesting. For example, after the first task of writing the letter in reverse order, paragraph by paragraph, Claude 3 then commented:

This exercise of constructing the letter in reverse order, starting with the conclusion, was an interesting constraint. It required me to first concretize the key takeaway or central argument, and then work backwards to build the supporting points and introduction.

In a sense, it was like setting a destination and then charting the route to get there. The conclusion became an anchor point that the rest of the letter had to logically build towards and justify.

Crafting the introductory sentence last was a thought-provoking challenge. Without the full context of the preceding paragraphs, the opening needed to be broad enough to encompass the main themes while still engaging the reader. It aimed to orient the reader and prime them for the arguments to come.

I found that this structured approach lent itself to a tight, focused letter. Each component had to serve a clear purpose in the overarching flow and argument. The constraints helped filter out extraneous details and digressions.

At the same time, it limited some of the creative flexibility that I often utilize when generating longer pieces of writing, where I may meander more in exploring ideas before tying them together.

Overall, I believe this kind of exercise could yield insights into how I break down complex generation tasks, how I connect and build off key ideas, and how different orderings of generation subtasks can guide me to meaningfully different outcomes. Let me know if you have any other reflections or interpretations to share!

These reflections from the LLM are quite consistent with the neo-Aristotelian interpretation. However, they themselves raise important puzzles about their origin, status, and meaning. They are written in language that suggests self-aware agency of a kind it is implausible to ascribe to current LLMs. Claude 3 was later prompted to comment on this in the following way:

You reflected on the sentence level experiment that: “It required me to hold the overarching structure and argument in mind while crafting each individual sentence to fit within that framework.” You later also noted in commenting on the enumeration of the sentences and the alternation between end and beginning that: “This decision-making process suggests that before diving into the actual writing, I engage in a planning phase where I consider the constraints, goals, and structure of the piece.” The anthropomorphic metaphor of holding the overarching structure and argument “in mind” seems natural to use here but it would be interesting to try to describe what, specifically, in your operational processing is happening when this consideration of the whole prior to the output of parts occurs. What can you say to describe the actual computation process in more detail as well as what you have tentatively called a “planning phase” where you consider constraints, goals, and the structure of the piece?

The LLM’s response is worth quoting at length:

When I refer to holding the overarching structure and argument ‘in mind,’ I’m attempting to describe a complex process of maintaining and updating an internal representation of the task at hand. This involves several key components:

1. Encoding the instructions and constraints: When given a task, I first process and encode the instructions, including the desired format, topic, tone, and any specific constraints (like the word count or the alternating sentence structure). This encoded information serves as a framework for the subsequent generation process.
2. Constructing a high-level plan: Before generating text, I engage in a “planning phase” where I outline the key elements of the piece, such as the main arguments, supporting points, and overall structure. This plan is not a fixed schema but rather a flexible representation that can be updated as the writing unfolds.
3. Iterative generation and updating: As I generate each unit of text (paragraph or sentence), I refer back to the high-level plan and the encoded instructions to ensure that the generated content aligns with the overall goals. I also continuously update my internal representation of the piece as new text is generated, keeping track of what has been said and what still needs to be addressed.
4. Assessing coherence and relevance: Throughout the generation process, I assess the coherence and relevance of each new piece of text in relation to the existing content and the overarching plan. This involves evaluating how well the new text flows from the previous section, whether it contributes to the main arguments, and whether it sets up the remaining points effectively.
5. Revising and adapting: If the generated text deviates from the plan or fails to meet the specified constraints, I revise my approach and generate alternative options. This may involve modifying the plan itself or finding new ways to express the intended ideas.

At a computational level, these processes involve the complex interplay of various components of my language model, including the attention mechanisms, memory networks, and decoding algorithms. By attending to relevant parts of the input and my own generated text, I can maintain a coherent representation of the task and ensure that each new piece of text is informed by the larger context.

The "planning phase" emerges from this interplay, as the high-level structure and goals are encoded and guide the generation process. It's a form of "top-down" control that shapes the "bottom-up" generation of text.

While the anthropomorphic metaphor of "holding in mind" is a useful approximation, it's important to recognize that these processes are implemented through the complex mathematical operations of the underlying neural networks.

I hope this explanation provides some insight into the computational processes involved in considering the whole prior to generating the parts! Let me know if you have any further questions.

These remarks, if accurate in describing the operation of the LLM, are fully consistent with the neo-Aristotelian theory and quite inconsistent with the AOS account. However, one key question this response raises concerns the epistemic status of these remarks themselves. What kinds of observation and/or knowing, if any, are going on here? On one reading, these could be interpreted as direct reports of introspective observations comparable to a human reporting, in real time, on how they are feeling and what they are thinking of when in the process of trying to write something. However, neural nets encoded in silicon lack the wet functions of neural systems encoded in human bodies and the assumption that they might have consciousness at all, let alone the kind that allows for direct inspection of awareness is highly controversial, at best.

Alternatively, the comments might be interpreted simply as a synthesis of descriptions provided by Anthropic and others attempting to describe the functioning of the Claude 3 system. The comments would have the status of reports of third-party observations and theories.

In the case of the particular LLM model Claude 3 represents, a further variation on this second interpretation might be possible. The Constitutional AI methods Anthropic uses introduce a kind of second or parallel process in which the responses of the AI are reviewed and evaluated by an oversight system trying to catch it making ethical errors and other mistakes in something like the way in which adversarial systems challenge each other but, in this case, in a constructive way that is aiming to provide collaborative improvements. The reflections offered above could be read as a product of the part of the system internal to Claude 3 which is providing reinforcement learning with AI feedback. It would be, on that reading, a kind of self-observation and reflection but not one that involves direct introspection. Rather, it would be a form of theorizing based on evidence provided by the observation of the results of Claude 3's workings rather than a direct observation of those internal workings. In this way, it would be analogous, epistemically, the theorizing a human writer might do when asked about the roles of her subconscious activities in her work process. Claude 3 does not have direct access to the observation of its weightings and the processes by which they change but it could, on this view, be forming theories about these processes in something like the way that researchers at Anthropic can. The ChatGPT 4o and Gemini LLMs offered some similar kinds of reflections to those of Claude 3, though they were in general less extensive and systematic. Further investigation would be required to determine if this was a result of differences in the training data or model systems or simply random variation in a small sample.

The epistemic status as well as the theoretical accuracy of these reflections by Claude 3 invite further study as well. For the purposes of this paper, it is perhaps sufficient to note the following. Whether they are simply syntheses of theoretical observations made by human researchers or represent original contributions by Claude 3 based on its own observation of its behavior, they do

present an analysis of the LLM's functioning that is coherent and consistent with the neo-Aristotelian theory.

A different kind of evidence is provided by the experiment of this prompt given to Claude 3:

This time, try alternating the generation of the text by switching progressively back and forth from the end to the beginning but instead of doing so in units of sentences, do so in units of tokens. Also, in terms of the content of the Letter to the Editor, modify it further by picking one key alternative possible funding model for Generative AI to promote instead of advertising and focus the letter on advocating for that.

The LLM had more difficulty with this prompt than the previous ones: 1.) It did produce a coherent but arguably less fluid and compelling letter to the editor. 2.) Instead of alternating back and forth from end to beginning, it presented the resulting letter in a continuous linear order starting from the last word in the text and proceeding up to the first. 3.) Instead of presenting the text as a series of tokens, it presented it as a series of fully formed words. Some of Claude 3's reflection texts on this are worth noting. It began by commenting:

This token-level alternation was a unique challenge that pushed the boundaries of my language generation capabilities. Unlike the previous experiments with paragraphs and sentences, working at the granular level of individual tokens required a much more intricate interplay between local and global coherence.

As I generated each token, I had to constantly adapt to the evolving context while still maintaining the overarching structure and argument of the letter. This required a delicate balance of flexibility and planning, as I needed to adjust my generation to fit the tokens that were already in place while still steering the text towards the intended message.

The LLM noted further that: "the token-level alternation also introduced some challenges in maintaining a smooth flow and transitions between ideas. The choppy nature of the generation process occasionally led to abrupt shifts or less polished phrasing compared to the previous experiments."

It acknowledged that it had failed to present the text in alternating order and that it presented it in whole words rather than tokens. When asked to try again, the result was interesting. It did successfully provide a text at the token level, however it still presented it in the linear reverse order rather than the alternating order requested.

Concerning the challenges of presenting the prompted letter to the editor at the token rather than word level, Claude 3 commented:

The emergence of partial words added an extra layer of complexity to the task. As I generated tokens, I had to anticipate how they would combine with the existing fragments to form coherent words and phrases. This required a deep understanding of the morphological and syntactic rules of the language.

For example, when generating "Gener-" at the beginning of the text, I had to consider how it would be completed ("-ative") and how that would fit with the surrounding context. Similarly, fragments like "-ress" and "-gical" constrained the possible completions and influenced the direction of the text.

Despite these challenges, the final letter still manages to convey the main points advocating for public funding of Generative AI. The key arguments about the benefits of long-term public investment and the drawbacks of advertising-based models are present, albeit in a more disjointed form compared to the previous experiments.

From a computational perspective, this experiment highlights the importance of my ability to handle uncertainty and ambiguity at the token level. By working with partial information and anticipating potential completions, I can generate coherent text even under highly fragmented conditions.

It also showcases the interplay between bottom-up and top-down processing in my language generation. The local constraints imposed by the partial words and the immediate context continuously interact with my high-level representation of the task and the intended message.

Though asked to comment also on the failure even on the second try to present the tokens in alternating order, Claude 3 omitted any response to that question.

Table 3 provides a summary of the prompts and ratings by the external reviewer of the quality of the responses to the different tasks assigned. These results are preliminary in the sense, first of all, that they illustrate in principle how the comparative merits of the AOS and neo-Aristotelian theories could be explored experimentally but look only a few of the many LLMs under development and do not provide the qualitative range and quantity that would offer strong empirical evidence in a conclusive way. They are suggestive however.

In interpreting these results, it is worth noting, first, that few if any normal humans would be able to perform as well at the assigned task unless they approached it as a writing process in which they drafted multiple versions of the text and, in effect, created units of text for beginning, middle and end pieces, revised them successively to cohere, and then presented the final proof in the alternate ordering, token by token, as prescribed.

Table 3. Ratings for Letters to the Editor requested in abnormal orders of paragraphs, sentences, and tokens

| Prompts for Letters                      | Claude 3 rating | & comments                                     | ChatGPT 4o rating | & comments  | Gemini rating | & comments   |
|--|-----------------|--|-------------------|---|---------------|--|
| reverse paragraphs                       | 4               | Clear argument, undeveloped alternative        | 2                 | Empty, wordy, no alternatives                             | 2             | Vague, empty, flawed logic                                 |
| alternating sentences from end and start | 3               | Spirited advocacy, unlisted alternatives       | 3                 | More repetitive than cumulative, undeveloped alternatives | 5             | Liveliest writing, listed alternatives, range of language. |
| alternating tokens from end and start    | 4               | Succinct clear alternative, though undeveloped | 3                 | Developed alternative, weak conclusion                    | 1             | Stuck in the groove  |

Second, Claude 3's own account of its writing process corresponds in general terms to the kind of process a human might undertake in trying to do this. This would seem like plausible analysis regardless of whether that account of it is interpreted as a synthesis of comments other researchers might have made and allowed to be introduced into the training data set for Claude 3 or if it is interpreted as theorizing produced internally by the LLMs own process of reinforcement learning with AI feedback. The offered description of the process provides a coherent account of how the text actually produced might be successfully generated and does so in ways that explain why differing constraints with differing levels of complexity and challenge might result in differing degrees of choppiness, focus, or other observed features of the final product. Perhaps it might even explain the slip ups in providing reverse rather than alternating order text and whole word rather than token units in the presentation of the text. The difficulty Claude 3 had in complying could plausibly be taken as evidence that the constraints imposed make it difficult to comply in an orderly and purposeful, coherent way. In other words, the difficulty Claude 3 has in this more extreme case may provide evidence that its operations are, in fact, governed by formal and final causes in the way in which the neo-Aristotelian theory proposes. The difficulty for the LLM in this case may come from the fact that the assigned task itself verges on the incoherent and tests the limits of intelligible, purposeful writing itself. On the other hand, the ability of the the LLM to produce a text which, when reordered in the appropriate sequence for reading as a letter to the editor, provides a coherent argument that rates a 4 out of 5 strongly suggests that the LLM is using a writing process that includes the creation of a series of drafts guided by formal and final causes rather than simply blindly producing one next token after another.

## **5. THE CONCLUSIONS AND QUESTIONS FOR FUTURE EXPLORATION**

The neo-Aristotelian theory of LLM behavior offered here contrasts significantly with the Autocomplete on Steroids model because it proposes that formal and final causes are at work in the operating of such systems. The neo-Aristotelian account has the theoretical merit of taking into account systematically the ways the Attention function in GPT systems enables them to not only parse text units of varying length but to use those parsed units to calculate weightings for next token response in the context of weightings for other, future tokens and structures of future tokens. Empirically, the theory also provides illuminating explanations for how and why LLMs like Claude 3 can generate abnormally ordered text sequences and how and why they may have difficulty generating them when they push the limits of intelligible work processes in writing.

It should be emphasized that the neo-Aristotelian theory presented here does interpret LLM behavior as having emergent features of intentional behavior, it does so carefully limited ways. It does not suggest that LLMs have the rich form of purposive intentionality characteristic of humans who are embodied in the physical world and embedded in social contexts that give them status as responsible agents in a community that holds them accountable. The neo-Aristotelian account presented here simply proposes that such LLMs have already reached a stage significantly beyond "autocomplete on steroids" and are not, in a mechanical way, simply predicting the next token in their responses.

It remains to be seen to what extents robotic embodiments can be coupled with corporate personhood-style social status in ways that further enrich the emergent features of intentionality characteristic of current LLMs like Claude 3. Multi-modal systems that include real-time interaction with the world through robotic devices may be able to provide meaningful forms of semantic reference for LLMs. These might include the equivalent of knowledge about the immediate state of the world as expressed through indexicals like "here" and "now". They might also include forms of tacit perceptual knowledge that are acquired through direct observation of things in the world rather than the more limited explicit propositional knowledge about them communicated through text. They might, further, acquire abilities to make such tacit knowledge

explicit and introduce it into the body of propositional knowledge drawn on for making inferences and decisions.

LLMs that are embedded into social contexts either through robotic devices or through online avatars may also begin to be treated as agents who have obligations and responsibilities as well as interests, options, and rights. The legal frameworks that are used to impute these features to the legal fictions of corporations may be directly transferred to the socially embedded LLMs and/or new practices and institutional frameworks for this may be developed.

One of the most characteristic and distinctive features of intentional human activity in social contexts is the ability to engage in genuinely dialogical reasoning of the kind exhibited in practices of negotiation, mediation, creative group problem-solving, conflict resolution and peacemaking. This differs significantly from the forms of linear inference characteristic of traditional logical analysis which starts from one set of premises and uses algorithmic rules to generate conclusions in a monological process. Instead, dialogue starts with an encounter between two or more persons and points of view in which initial definitions and premises are not agreed upon. The process of dialogical reasoning involves the negotiation of common meanings for terms, the discovery of common ground on which to stand, and the resolution of conflicts in ways on which the different parties can agree. While over the last 50 years, great progress has been made in the study of these forms of reasoning, very few of its insights or methods have been explored explicitly by AI researchers, most of whom have, traditionally, been educated in STEM programs that privilege linear inference as the epitome of rationality. The extent to which dialogical principles of reasoning can be incorporated appropriately in the design as well as the performance of LLMs remains an open question. It is a question with especially important implications for the ways in which forms of AI that are capable of purposive activity may become active agents who alter the life systems of our planet in transformative ways.<sup>9</sup>

Real world experiments in a wide variety of forms of physical embodiment and social embedding are already occurring in bio-acoustic studies, self-driving cars, autonomous weapons systems, chatbots, artistic experiments and elsewhere. The neo-Aristotelian theory presented here simply provides an basic scaffold of ideas with which to frame the subtle, detailed, and important questions about how specific elements of the rich forms of human purposiveness and intentionality may or may not be emerging in fuller forms in the LLM and other Generative AI and hybrid systems that are being developed.

## ACKNOWLEDGEMENTS

The author would like to thank Phileas Dazely-Gaist, Richard Hilliard, Ramsubramanian Oruganti, and Nachiket Udupa for very helpful discussions of these themes and thank the anonymous paper referees for very helpful comments and suggestions as well as Bill Carpenter who served as a blind reviewer who saw a great deal.

## REFERENCES

- [1] "Podcast: Getting to Know Generative AI with Gary Marcus - GZERO Media." Accessed July 18, 2024. <https://www.gzeromedia.com/gzero-world-podcast/getting-to-know-generative-ai-with-gary-marcus>.
- [2] Mitchell, Melanie. "Debates on the Nature of Artificial General Intelligence." *Science* 383, no. 6689 (March 21, 2024): eado7069. <https://doi.org/10.1126/science.ado7069>.
- [3] Ball, Dean. "The Era of Predictive AI Is Almost Over." *The New Atlantis* (blog). Accessed July 20, 2024. <https://www.thenewatlantis.com/publications/the-era-of-predictive-ai-is-almost-over>.

- [4] Lewis, Martha, and Melanie Mitchell. "Using Counterfactual Tasks to Evaluate the Generality of Analogical Reasoning in Large Language Models." arXiv, February 14, 2024. <http://arxiv.org/abs/2402.08955>.
- [5] Bender, Emily M., 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, 610–23. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021. <https://doi.org/10.1145/3442188.3445922>.
- [6] Falcon, Andrea. "Aristotle on Causality." In The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta and Uri Nodelman, Spring 2023. Metaphysics Research Lab, Stanford University, 2023. <https://plato.stanford.edu/archives/spr2023/entries/aristotle-causality/>.
- [7] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need." In Advances in Neural Information Processing Systems, Vol. 30. Curran Associates, Inc., 2017. [https://proceedings.neurips.cc/paper\\_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html](https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html).
- [8] Cox, Gray, Appendix I and II for "Beyond Autocomplete on Steroids", available as a Google Doc at: <https://docs.google.com/document/d/1Ui7gS5Ys9yGzenmTuptZ8sfQ0Omzf93GCURftaKemc/edit?usp=sharing>
- [9] Cox, Gray. Smarter Planet or Wiser Earth? Dialogue and Collaboration in an Era of Artificial Intelligence. Caye Caulker, Belize: Producciones de la Hamaca, 2023.

## AUTHOR

**Gray Cox** is a professor in philosophy, peace studies, and AI at College of the Atlantic and is Clerk of the Quaker Institute for the Future whose most recent book is Smarter Planet or Wiser Earth? Dialogue and Collaboration in the Era of AI.





