Social sentience in neural language models

Alessandro Acciai[1,†], Pietro Perconti[1,†] and Alessio Plebe[1,†]

[1]University of Messina, Department Of Cognitive Science

Abstract

This work explore the ability of Neural Language Models (NLMs) to produce and modulate "autobiographical" stories, , thanks to their extensive exposure to social linguistic interactions, with a level of narrative coherence comparable to that of humans. Generative AI based on transformer architecture has demonstrated the ability to perform extraordinary tasks often considered exclusive to human cognitive abilities. The need to clarify the functioning of the algorithmic black box within transformers, combined with the opportunity to use cognitive science tasks and tests in this investigation, has led to a significant field of studies aiming to bridge this explanatory gap. The term "machine psychology" refers to the administration of cognitive tests, typical of human cognition, to NLMs. Contributing to this debate our proposal involves an empirical study on the modulation of autobiographical narrative coherence, an element widely used in cognitive psychology for studying aspects related to self-integrity and fragmentation, emotion modulation, worldview and self-construction. We subjected OpenAI models to tasks requiring story production following a multi-level pre-induction framework, considering three variables: age, mood, and gender. The results demonstrate that NLMs are not only capable of simulating various aspects of the human experience but can also adapt to the designated role and modulate their level of narrative coherence accordingly. This provides evidence of these artificial artifacts' ability to produce cognitively complex textual elaborations and suggests that the emergence of narrative awareness within transformer architecture, akin to the prelude to consciousness in human, may be possible due to their overexposure to social linguistic interactions.

Keywords

Neural Language Model, Artificial Sentience, Narrative coherence

1. Introduction

The current Neural Language Models (NLMs), derived from the successful invention of the Transformer architecture [1], represent a peculiar and unusual type of entity, even from a scientific research standpoint. They are the only non-biological entities capable of cognitive performances that, in many respects, are surprisingly close to human ones. At the same time, they are man-made objects, but their design does not clarify on how their range of cognitive abilities is realized. Therefore, they require a search for explanations, not unlike the research typically required by complex natural systems.

The Transformer model fundamentally represents a system that ensures highly efficient textual processing by capturing the relationships between words within the produced and required text. Its structure, based on simple linear algebra, allowed for overcoming the challenges faced by earlier ANN-based systems. Firstly, it transforms words into vectors through word embedding [2], significantly simplifying the manipulation of the semantic aspects of language. Secondly, the introduction of the attention mechanism [3] allows for all words to be vectorized and presented simultaneously as input to the architecture, which can track all relationships between each word within the processing. Finally, the autoencoder mechanism addressed the problem of supervised learning by borrowing the autoencoder technique [4], where the input task is reproduced in the output, effectively aligning the encoder and decoder.

Even though there is no underlying claim to simulate human cognitive functions, and without any specific training in this regard, Transformer-based Neural Language Models have demonstrated abilities that go far beyond translation and simple language processing [5]. The numerous similarities

AIxPAC-2nd Workshop on AI for Perception and Artificial Consciousness - AIxIA '24, November 25-28, 2024, Bozen, Italy

*Alessandro Acciai

†These authors contributed equally.

alessandro.acciai@studenti.unime.it (A. Acciai); pietro.perconti@unime.it (P. Perconti); alessio.plebe@unime.it (A. Plebe)

0000-0002-2547-9113 (A. Acciai); 0000-0002-3633-098X (P. Perconti); 0000-0003-3666-061X (A. Plebe)

of NLMs with human cognitive performance in various cognitive tests and the lack of understanding of the mechanisms capable of supporting it have suggested applying the methods of sciences that have traditionally focused on the mind psychology and cognitive science to them. This proposal has been named "machine psychology" [6], and it has quickly produced various important results [7, 8].

This line of research has shown that the processing capabilities of NLMs cannot be explained solely through the word prediction function [9, 10, 11, 12] and that, even through massive exposure to large linguistic corpora, they can complete cognitive tasks that have so far been exclusive to humans [13].

This work fits into this line of research, exploring this capability and arguing that it is not just exposure to linguistic corpora but also the significant presence of social linguistic interactions within them that form one of the bases of the abilities demonstrated by NLMs.

To support this hypothesis, our study utilizes a specific analysis of narrative production by human subjects, aiming to outline certain mental characteristics based on the coherence traceable in the text. Specifically, we will employ a well-established psychological analysis scheme, known as the Narrative Coherence Coding Scheme (NaCCS) [14], to analyze the modulation of coherence in the stories produced by the OpenAI family of NLMs. Specifically, we prompted GPT-3 and GPT-4 to produce texts with three variable prompts, thereby inducing variations in age, mood, and gender.

In the first part of this work, we will explore the importance of narrative construction in the creation of self identity through its social character and how this role is important in the study of consciousness. In the second part, we will illustrate our experimental study on narrative coherence in NLMs and in light of the results and the literature examined, we will conclude with reflections on whether the abilities demonstrated by NLMs suggest the presence of narrative awareness, a primordial form of consciousness in human beings.

2. Machine Consciousness from a Social-Narrative Perspective

The attempt to understand consciousness within a scientific framework is a relatively recent endeavor, only a few decades old. While some theories of consciousness, grounded in neuroscience and computational model, such as the Global Neuronal Workspace Theory and Integrated Information Theory, have drawn considerable attention within the consciousness research community, there remains no universally accepted framework, and even the search for the neural correlates of consciousness has yet to yield conclusive results [15]. Nevertheless, research into machine consciousness has continued to develop [16, 17]. More recently, leveraging techniques derived from Natural Language Models (NLMs), researchers have begun exploring whether deep learning-based cognitive architectures can offer promising results in the realm of consciousness, as they have in language processing [18, 19]. It is still too early to draw definitive conclusions about this research direction, but one characteristic stands out as particularly interesting for the purposes of this paper.

It seems that several models of artificial consciousness are socially oriented. This means that the self-awareness we aim to model in machines appears to serve primarily social purposes. Consider, for example, studies aimed at modeling inner speech in humanoid robots [20]. Although this capability can improve conscious performance in tasks that are not directly related to sociality, such as passing the Mirror Test [21], inner speech in humanoid robots generally appears to create an internal logical space where the social consequences of various possible actions can be simulated offline before one is chosen and executed. Observing this type of behavior supports the social hypothesis of self-consciousness, which proposes that self- consciousness primarily serves social cognition purposes [22, 23]. The ability to represent oneself as a character in one's own life is a very common and natural way of exercising self-consciousness and situating the individual within a real or imagined social network. In other words, narrativity and its social character is a key component of self-consciousness. However, it is only one of the ways in which self-consciousness happens and contributes to shaping one's singular personality, alongside episodic and sentimental personality types [24]. Investigating how the capacity to construct narratives plays a central role in the stream of consciousness and reflexive reasoning is crucial for advancing machine consciousness. This is why, in the spirit of machine psychology and

with the conviction that testing what we think we know about the human mind on machines, and vice versa—applying to humans what we learn from machines—is the best way to advance cognitive science as a whole. This is precisely what we aim to do, and we will describe it in the following section.

3. Narrative coherence in NLMs

Coherence serves as a measuring tool for various significant aspects of our personal narrative [25], and the way we construct and reconstruct our experiences influences the meaning we attribute to events in our personal life [26]. For example, it has been shown that the extent to which individuals coherently narrate their autobiographical memories is related to their mental health [27, 28, 29]. Narrative coherence is indeed of great help in the psychological analysis of a subject and can be defined as the extent to which life narratives (global coherence) [30, 31, 32, 33] or narratives of a single event (local coherence) [34, 14] make sense to a naive listener and are able to convey the content and meaning of the described events in a structurally and thematically coherent manner.

The process of constructing a narrative allows individuals to derive meaning from their lived personal experiences and influences the regulation of emotions associated with them. Therefore, the way people talk about key events in their lives reflects their emotional adaptation [35, 36], influencing psychological well-being [37, 38]. Many studies have shown that higher narrative coherence is associated with lower internalizing symptoms and greater psychological well-being [39, 40, 41]. Furthermore, cross-sectional studies demonstrate that individuals whose personal narratives exhibit high narrative coherence have lower levels of psychopathology [42].

According to Reese [14], narrative coherence cannot be a singular construction but must emerge from multidimensional aspects that contribute to the overall narrative from various focal points, independently of each other, and develop at varying rates across the lifespan. Reese's proposal for assessing coherence thus includes three independent dimensions that are influenced by different developmental factors across the lifespan, as outlined in a three-factor rating grid with: Context (narrative more or less defined in terms of space and time); Chronology (linearity of the logical and chronological structure); Theme (emotional elaboration, resolution, closure, a connection to other important events, or the self). The sum of these three dimensions, according to a scoring scale from 0 to 3, gives the global coherence score of the narrative.

The stories were generated by GPT-3.5 and GPT-4 according to a pre-dialogue with prompt induction on 3 variables: Age, Mood, and Gender.

- **Age**: For the age variable, four age groups were simulated similarly to Reese's study: Child 3 to 11 years, Teenage 12-14 years, Midlife 20 to 36 years, Adult 52 years;

- **Mood**: Regarding the aspect of emotion modulation, the NLMs were asked to narrate a particularly positive event (Positive), a negative event (Negative), or no specific guidance was introduced (Neutral);

- **Gender**: The stories were balanced by gender, with half narrated by male and half by female characters.

The main results of the analyses conducted on narrative coherence and its individual dimensions for stories generated by GPT-3.5 turbo, GPT-4, and the average obtained from both models indicate that age, gender, and mood can differently influence the narrative coherence of stories generated by GPT-3.5 and GPT-4, as shown in the table. The models show significant results both collectively and individually, further confirming better performance in GPT-4. By comparing the narrative production of the two models, taking into account the trends in coherence dimensions concerning age, we obtained interesting results for both NLMs. Both models exhibited a similar downward trend in overall coherence scores and individual coherence dimensions across different age groups, maintaining good levels of coherence despite some deterioration in older age groups. Overall, GPT-4's narrative production was

richer and more coherent across all age groups compared to GPT-3.5, confirming the superiority of OpenAI's larger model.

The emotional induction reveals particularly interesting data. Specifically, the study shows that inducing a specific mood, whether positive or negative, positively influences the coherence trend. For both models, the request to narrate particularly negative events had the greatest impact on overall coherence and on the Theme dimension, significantly increasing them, with more pronounced results in GPT-4. Finally, the data revealed that no significant differences were found concerning the induction of gender differences.

The overall results demonstrate the good level of multidimensional development of narrative coherence in the NLMs examined and confirm that the textual production of GPT-3.5 and GPT-4 is not only formally correct but also narratively very coherent, achieving results similar to or even superior to those found in studies with human samples [14].

The autobiographical narrative productions developed along the multidimensional trajectory of the NaCCS are thus very on-topic with respect to the subject matter, providing precise temporal and spatial references, unfolding along a timeline that, even if not always explicitly defined, is precise and in line with the narrated event. As we will see in detail, the results align with several studies on NLMs, demonstrating the ability of the Transformer architecture to simulate cognitive functions that, in humans, require the activation of very complex mechanisms.

The results of our work add to this picture. The consistency demonstrated in autobiographical narratives generated by the models is far from trivial, and if in human beings it denotes a fundamental integrity of the personal self, it is valid to hypothesize that something similar, albeit with the necessary differences and limitations, is being constructed in language models.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 6000–6010.

[2] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.

[3] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: International Conference on Learning Representations, 2016.

[4] G. Hinton, R. S. Zemel, Autoencoders, minimum description length and Helmholtz free energy, in: Advances in Neural Information Processing Systems, 1994, pp. 3–10.

[5] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, Y. Zhang, Sparks of artificial general intelligence: Early experiments with GPT-4, arXiv abs/2303.12712 (2023).

[6] T. Hagendorff, Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods, arXiv abs/2303.13988 (2023).

[7] M. Binz, E. Schulz, Using cognitive psychology to understand GPT-3, Proceedings of the Natural Academy of Science USA 120 (2023) e2218523120.

[8] M. Kosinski, Theory of mind may have spontaneously emerged in large language models, arXiv abs/2302.02083 (2023).

[9] L. Floridi, GPT-3: Its nature, scope, limits, and consequences, Minds and Machines 30 (2020) 681–694.

[10] E. M. Bender, T. Gebru, A. McMillan-Major, S. 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, ACM, 2021, pp. 610–623.

[11] M. W. Eysenck, C. Eysenck, AI vs Humans, Routledge, Abingdon (UK); New York, 2022.

[12] L. Miracchi Titus, Does ChatGPT have semantic understanding? a problem with the statistics-of-occurrence strategy, Cognitive Systems Research 82 (2024) 101174.
[13] S. Trott, C. Jones, T. Chang, J. Michaelov, B. Bergen, Do large language models know what humans know?, Cognitive Science 47 (2023) e13309.
[14] E. Reese, C. A. Haden, L. Baker-Ward, P. Bauer, R. Fivush, P. A. Ornstein, Coherence of personal narratives across the lifespan: A multidimensional model and coding method, Journal of Cognition and Development 12 (2011) 424–462.
[15] A. K. Seth, T. Bayne, Theories of consciousness, Nature Reviews Neuroscience 23 (2022) 439–452.
[16] I. Aleksander, The world in my mind, my mind in the world, Andrews UK Limited, 2013.
[17] T. Bayne, A. K. Seth, M. Massimini, J. Shepherd, A. Cleeremans, S. M. Fleming, et al., Tests for consciousness in humans and beyond, Trends in Cognitive Sciences (2024).
[18] P. Butlin, R. Long, E. Elmoznino, Y. Bengio, J. Birch, A. Constant, G. Deane, S. M. Fleming, C. Frith, X. Ji, R. Kanai, C. Klein, G. Lindsay, M. M. L. Mudrik, M. A. K. Peters, E. Schwitzgebel, J. Simon, R. VanRullen, Consciousness in artificial intelligence: Insights from the science of consciousness, arXiv abs/2308.08708 (2023).
[19] D. J. Chalmers, Could a large language model be conscious?, arXiv (2024). URL: https://arxiv.org/abs/2303.07103.
[20] A. Chella, A. Pipitone, A. Morin, F. Racy, Developing self-awareness in robots via inner speech, Frontiers in Robotics and AI 7 (2020) article 16.
[21] A. Pipitone, A. Chella, Robot passes the mirror test by inner speech, Robotics and Autonomous Systems 144 (2021) 103838.
[22] P. Perconti, Rethinking subjectivity: The social roots of consciousness, Epistemology and Philosophy of Science (2024). In press.
[23] A. Plebe, P. Perconti, The Future of the Artificial Mind, CRC Press, Boca Raton, 2022.
[24] P. Perconti, Identity, narratives and psychopathology: A critical perspective, in: V. Cardella, A. Gangemi (Eds.), Psychopathology and The Mind. What mental disorders can tell us about our minds, Routledge, London, 2021, pp. 215–221.
[25] J. S. Bruner, Acts of meaning: Four lectures on mind and culture, Harvard University Press, Cambridge (MA), 1990.
[26] C. Linde, Life stories: The creation of coherence, Oxford University Press, Oxford (UK), 1993.
[27] Y. Chen, H. McAnally, Q. Wang, E. Reese, The coherence of critical event narratives and adolescents' psychological functioning, Memory 20 (2012) 667–681.
[28] K. C. McLean, A. V. Breen, M. A. Fournier, Constructing the self in early, middle, and late adolescent boys: Narrative identity, individuation, and well-being, Journal of Research on Adolescence 20 (2010) 166–187.
[29] E. Reese, E. Myftari, H. M. McAnally, Y. Chen, T. Neha, Q. Wang, F. Jack, S. Robertson, Telling the tale and living well: Adolescent narrative identity, personality traits, and well-being across cultures, Child Development 88 (2017) 612–628.
[30] T. Habermas, S. Bluck, Getting a life: the emergence of the life story in adolescence, Psychological Bulletin 126 (2000) 748.
[31] T. Habermas, C. de Silveira, The development of global coherence in life narratives across adolescence: temporal, causal, and thematic aspects, Developmental Psychology 44 (2008) 707.
[32] T. Habermas, E. Reese, Getting a life takes time: The development of the life story in adolescence, its precursors and consequences, Human Development 58 (2015) 172–201.
[33] C. Köber, F. Schmiedek, T. Habermas, Characterizing lifespan development of three aspects of coherence in life narratives: a cohort-sequential study, Developmental Psychology 51 (2015) 260.
[34] D. R. Baerger, D. P. McAdams, Life story coherence and its relation to psychological well-being, Narrative Inquiry 9 (1999) 69–96.
[35] J. M. Adler, T. E. Waters, J. Poh, S. Seitz, The nature of narrative coherence: An empirical approach, Journal of Research in Personality 74 (2018) 30–34.
[36] T. E. A. Waters, C. Köber, K. L. Raby, T. Habermas, R. Fivush, Consistency and stability of narrative coherence: An examination of personal narrative as a domain of adult personality, Journal of

Personality 87 (2017) 151–162.
[37] S. N. Haber, Neural circuits of reward and decision making: Integrative networks across corticobasal ganglia loops, in: R. B. Mars, J. Sallet, M. F. S. Rushworth, N. Yeung (Eds.), Neural Basis of Motivational and Cognitive Control, MIT Press, Cambridge (MA), 2011, pp. 22–35.
[38] J. L. Pals, Narrative identity processing of difficult life experiences: Pathways of personality development and positive self-transformation in adulthood, Journal of Personality 74 (2006) 1079–1110.
[39] J. P. Lilgendahl, D. P. McAdams, Constructing stories of self-growth: How individual differences in patterns of autobiographical reasoning relate to well-being in midlife, Journal of Personality 79 (2011) 391–428.
[40] E. Vanderveren, P. Bijttebier, D. Hermans, Autobiographical memory coherence and specificity: Examining their reciprocal relation and their associations with internalizing symptoms and rumination, Behaviour Research and Therapy 116 (2019) 30–35.
[41] T. E. A. Waters, R. Fivush, Relations between narrative coherence, identity, and psychological well-being in emerging adulthood, Journal of Personality 83 (2015) 441–451.
[42] M. Lind, S. Vanwoerden, F. Penner, C. Sharp, Inpatient adolescents with borderline personality disorder features: Identity diffusion and narrative incoherence, Personality Disorders: Theory, Research, and Treatment 10 (2019) 389.