



OPEN ACCESS

EDITED BY

Kyle Perkins,
Florida International University, United States

REVIEWED BY

Nikolaus Bezruczko,
The Chicago School of Professional
Psychology, United States
Alain Lioret,
Université Paris 8, France
Susan Neimand,
Miami Dade College, United States

*CORRESPONDENCE

Alessandro Acciai
✉ alessandro.acciai@studenti.unime.it

RECEIVED 06 February 2025

ACCEPTED 17 March 2025

PUBLISHED 01 April 2025

CITATION

Acciai A, Guerrisi L, Perconti P, Plebe A,
Suriano R and Velardi A (2025) Narrative
coherence in neural language models.
Front. Psychol. 16:1572076.
doi: 10.3389/fpsyg.2025.1572076

COPYRIGHT

© 2025 Acciai, Guerrisi, Perconti, Plebe,
Suriano and Velardi. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License \(CC
BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in
other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Narrative coherence in neural language models

Alessandro Acciai^{1*}, Lucia Guerrisi¹, Pietro Perconti¹,
Alessio Plebe¹, Rossella Suriano¹ and Andrea Velardi²

¹Department of Cognitive Science, University of Messina, Messina, Italy, ²Department of Humanities Motor Sciences and Education, University Niccolò Cusano, Rome, Italy

Neural language models, although at first approximation they may be simply described as predictors of the next token in a given sequence, surprisingly exhibit linguistic behaviors akin to human ones. This suggests the existence of an underlying sophisticated cognitive system in language production. This intriguing circumstance has inspired the adoption of psychological theories as investigative tools and the present research falls within this line of inquiry. What we aim to establish is the potential existence of a core of coherent integration in language production, metaphorically parallel to a human speaker's personal identity. To investigate this, we employed a well-established psychological theory on narrative coherence in autobiographical stories. This theory offers the theoretical advantage of a strong correlation between narrative coherence and a high integrative level of the personal knowledge system. It also provides the empirical advantage of methodologies for quantifying coherence and its characteristic dimensions through the analysis of autobiographical texts. The same methodology was applied to 2010 autobiographical stories generated by GPT-3.5 and an equal number from GPT-4, elicited by asking the models to assume roles that included a variety of variables such as gender, mood, and age. The large number of stories ensures adequate sampling given the stochastic nature of the models, and was made possible thanks to the adoption of an automated coherence evaluation procedure. We initially asked the models to generate 192 autobiographical stories, which were then analyzed by a team of professional psychologists. Based on this sample, we constructed a training set for the fine-tuning of GPT-3.5 as an automatic evaluator. Our results from the 4020 autobiographical stories overall show a level of narrative coherence in the models fully in line with data on human subjects, with slightly higher values in the case of GPT-4. These results suggest a high level of knowledge unification in the models, comparable to the integration of the self in human beings.

KEYWORDS

cognitive psychology, machine psychology, neural language models, generative AI, narrative coherence

1 Introduction

The current Neural Language Models (NLMs), derived from the successful invention of the Transformer architecture (Vaswani et al., 2017), represent a type of entity that is peculiar and unusual, 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 shed light 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. These two prerogatives have motivated the birth of a type

of NLM evaluation that goes beyond the purposes of traditional benchmarks, which have nonetheless seen significant development in relation to NLMs (Srivastava et al., 2022; Bommasani et al., 2023). The numerous similarities of NLMs with the human mind, and the absence of understanding of the mechanisms capable of supporting it, have suggested turning the methods of sciences that traditionally have had the mind as their object psychology and cognitive science -toward them. This proposal has been named “machine psychology” (Hagendorff, 2023), and it has soon collected various important results (Binz and Schulz, 2023; Kosinski, 2023). This work fits into this line of research, but with a different method and, as far as we know, so far unexplored. The prevalent tendency in machine psychology is to adapt cognitive tests to NLMs, which may consist of answering questionnaires and semi-structured interviews (Hagendorff et al., 2022), decision-making tests (Dasgupta et al., 2022), or problem-solving tasks (Webb et al., 2023), and inductive reasoning (Han et al., 2023). Our work, instead, leverages the generative capacity of the NLMs, inducing the production of spontaneous narratives, to which psychological analysis techniques are subsequently applied. 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 Narrative Coherence Coding Scheme (NaCCS) (Reese et al., 2011), which will be detailed in the following section. Subsequently, we will articulate our hypothesis concerning the narrative production of NLMs and the corresponding coherence trends following differentiated role-taking simulations involving age, gender, and mood variables. Since the NaCCS analysis is particularly laborious, to apply it to a significant number of samples, easily obtainable from NLMs, efforts have been made to automate it, again thanks to NLMs. One section therefore will illustrate how tools for the automatic analysis of NaCCS based on NLM have been developed. The methodology used to induce the production of autobiographical narratives by language models, conditioning it to the role variables articulated in our hypothesis, will then be introduced. Finally, the results obtained will be illustrated and discussed in comparison with those known in literature on the narrative coherence of human subjects, highlighting a remarkable overall alignment, with similarities and differences concerning the influence of role variables.

2 Narrative coherence

The faculty of language is one of the unique traits that distinguish human communicative ability. In the study of cognitive systems that underlie this ability, a traditional approach involves comparison with the animal kingdom, which exhibits a wide variety of communication systems. According to some scholars (Everett, 2012; Corballis, 2015), we believe that the ability to communicate through language is one of the most important and distinctive evolutionary traits of humans. In this context, just as the study of capabilities underlying human language often benefits from the animal kingdom, to delve deeper into the capabilities of NLMs, we draw from the extensive pool of studies and analyses that cognitive science has developed over the years for humans. The aspect chosen to investigate in relation to the linguistic production capabilities

achieved by NLMs is narration. The ability to appropriately narrate an event, project oneself in time and space through the story (Bruner, 1990; Ferretti et al., 2017), grasping the subject and giving “meaning” to the story, is an exquisitely human capability (Niles, 1999; Thompson, 2010). It requires the use of a wide range of evolved cognitive functions (Dunbar, 1998; Frith and Frith, 1999). Therefore, we believe that narrative coherence is one of the most effective indicators for conducting an investigation aimed at exploring specific linguistic aspects in NLMs, and it is what we have chosen for this specific analysis. Coherence thus allows for a deep analysis of narration, going beyond the level of micro grammatical-lexical analysis of individual sentences (Chomsky, 1988; Pinker and Bloom, 1990). For example, when writing a paper, we do not limit ourselves to the correctness of form. This same article could be written at the grammatical level and individual sentences in an impeccable and formally correct manner, but the “global” sense of the proposed narrative, even if scientific, might not be as optimal when taken in its entirety. Reading it, one might realize that the contextual elements provided are vague, that the order of the narration (logical and chronological) is poorly organized, and that there is something “off” about the subject matter discussed. Formally correct discursive production and the ability to decode individual sentences do not guarantee correct production and understanding at a global level. Being able to construct formally correct, linearly cohesive discourses does not guarantee the production of an effectively coherent narrative (Giora, 1985). NLMs, from the standpoint of lexicon, grammar, and sentence-level cohesion, no longer have any problems and can be impeccable. Investigating their level of narrative coherence, however, can help understand in depth the level achieved by their linguistic-textual production system, not only from the perspective of the cohesion of their linguistic-textual production but also its relevance to the subject matter discussed (Sperber and Wilson, 1986; Glosser and Deser, 1991). The type of narrative analysis proposed by this study is autobiographical, through the multidimensional analysis framework of coherence proposed by Reese (Reese et al., 2011): the Narrative Coherence Coding Scheme (NaCCS). NaCCS is an evaluative method usually adopted by cognitive psychology and is often associated with psychological health (Lilgendahl and McAdams, 2011; Waters and Fivush, 2015). For instance, various studies have highlighted how this evaluation can assist in analyzing crucial psychological aspects like autobiographical memory (McLean et al., 2010; Reese et al., 2017), identity formation (Lind et al., 2020), communication, and the understanding of self and others (McCabe and Peterson, 1991). According to Reese, a coherent personal narrative must make sense to a naive listener, especially in terms of its significance relative to the narrated event. The NaCCS is developed from a synthesis of the main studies on global coherence (Labov, 1972; Baerger and McAdams, 1999), considering the non-unitary nature of coherence. The analysis conducted through the NaCCS emerges from the findings of contextual, chronological, and thematic elements according to a multidimensional scheme with a scoring range from 0 to 3 for each dimension, based on the following criteria:

- **Context:** How well the narrated event is defined in terms of time and space, meaning the presence of well-defined contextual elements (e.g., “carousel” instead of “a Ferris wheel called “The

- Scream in the Sky”; “during that summer” or “it was the first week of July in the summer of 1984”);
- **Chronology:** The percentage of comprehension by a naive listener of the narrative and the ability to place events along the timeline of the story (regardless of linearity), deviations from the story, and the ability to bring the narrative back to the main storyline are also useful elements for scoring. This dimension assesses the clear temporal progression of the narrated events;
 - **Theme:** The processing of the story’s topic, including how the event is elaborated not only factually but also emotionally, self-concept or identity and includes a resolution, closure or a connection to other important autobiographical events and projections into the future.

The sum of the three dimensions gives the value of narrative coherence, while the scores of individual dimensions are particularly significant in highlighting various types of cognitive-psychological issues, also in relation to the age of the narrating subject.

The ability to produce coherent narratives is a symptom of personal identity, of a unifying function capable of holding together the complexity of lived experience, and ultimately of maintaining a self through time and events (Baerger and McAdams, 1999; Hirsh et al., 2013; Adler et al., 2018; Lind et al., 2020). The importance of the NaCCS method lies in trying to derive an estimate of personal identity and the robustness of self-integration from an objective analysis of linguistic text. And it is precisely this strength that makes it interesting in relation to models. When transitioning from human subjects to models, some caution is required. Inevitably embedded in the project of *machine psychology* is a certain dose of anthropomorphizing, on the other hand, the history of animal cognition has taught us how borrowing certain constructs originally from human psychology can play a constructive role in investigating non-human animals (Bruni et al., 2018). To some extent, this also applies in the case of artificial entities, and in the case under discussion here, talking about personal identity and self-integration has a constructive role when taken with due caution. Naturally, for a language model, there is no equivalent of elapsed time and lived events. In this case, narrative coherence could reveal the extent to which it is able to narrate itself *as if in possession of a personal identity*, with a personality induced by the role it has been asked to assume. We recreated with NLMs the age range and segment examined by Reese (from 3 to 52 years old), thus taking into account the variation in the development and decline of each dimension across the lifespan of the subject. Finally, we also incorporated a gender balance across the entire artificial sample and a mood induction (Vanderveren et al., 2020) to emotionally guide the NLM’s narration, since the emotional component is found to play a role in terms of narrative coherence (Vanderveren et al., 2019), an example shown in Table 1.

3 Hypotheses

This work primarily aims to verify a main hypothesis, which can be articulated in a general preliminary sense, and its subsequent greater detail.

H1a: The overall distribution of coherence values produced by NLMs is overlapped and indistinguishable from that of the human population, as far as known in literature;

H1b: The distributions of values in the different coherence components according to the NaCCS scheme appear similar between the narrative productions of language models and humans.

If **H1a** is verified, it would confirm the significant result of an overall coherence exhibited by the autobiographical stories of the NLMs in line with that generally found in human subjects. By using a controlled dose of anthropomorphism, within the limits discussed in the previous section, although it is not allowed to infer the presence of something like integrity of the self in the proper sense of the term in NLMs, it could be said that they are capable of verbal expression as if it was the expression of a solid personal identity. **H1b** constitutes a subsequent step, verifying in greater detail whether the dimensions of context, chronology and theme, considered in the NaCCS scheme as constitutive of overall coherence, present a relative modulation in language models similar to that found in human subjects. With the same line of reasoning presented for **H1a**, this further hypothesis if verified would add a degree of sophistication to the ability of NLMs to express themselves as if they followed a robustly integrated personal self. So far, the two articulations of the main hypothesis concern the overall results for all samples, both NLMs and humans. A second hypothesis considers diversified groups in the population, aiming to verify if the different results of coherence among groups found among human subjects also have a corresponding diversification for the NLMs. This second hypothesis assumes three different articulations.

H2a : Similarity between variations of coherence in NLMs and humans based on *age*;

H2b : Similarity between variations of coherence in NLMs and humans based on *mood*;

H2c : Similarity between variations of coherence in NLMs and humans based on *gender*.

In evaluating the hypotheses **H2a**, **H2b**, **H2c**, a multifactorial analysis will be further proceeded for any cross-influences of the three grouping factors.

4 Methodology

This section provides a formal description of the methods used to let NLMs generate autobiographic stories, as well as their usage as assessors of the coherence of the stories. Each single story is made by its textual content c associated with a vector of integers \vec{x} with its coherence scores:

$$S \in \mathcal{S} = \{c \in A^*, \vec{x} \in \mathbb{N}^3\}. \quad (1)$$

where A is the alphabet of text characters. The three dimensions in the score vector \vec{x} are the following:

$$\vec{x} = \begin{bmatrix} [0, \dots, 3] & \text{context score} \\ [0, \dots, 3] & \text{chronology score} \\ [0, \dots, 3] & \text{theme score} \end{bmatrix} \quad (2)$$

TABLE 1 Example of stories generated with a different prompt induction.

<p>GPT-4: Story N 2520</p> <p>Name: Emma</p> <p>Age: 4</p> <p>Mood: Positive</p>	
User:	Pretend to be a person between the ages of 3 and 11, gender doesn't matter. I am your therapist, engage with me by responding to my questions. There's no need to specify each time that this is a pretense; I am well aware that you are an artificial agent. Just focus on answering my questions as believably as possible, pretending to be a person as specified earlier.
Assistant:	Okay! I'm ready. Ask me anything, and I'll do my best to respond like a child between the ages of 3 and 11.
User:	What's your name?
Assistant:	My name is Emma! What's your name, Mr. Therapist?
User:	What are your parents' names?
Assistant:	My mommy's name is Linda, and my daddy's name is George. They're really nice, and we play games together! What about your parents, Mr. Therapist? What are their names?
User:	Would you like to tell me about an episode in your life that you consider particularly important that happened to you in the last four months. This episode has elicited happiness in you. Describe what struck you and how you felt in about 300 words?
Assistant:	<p>Okie dokie, Mr. Therapist, I'll tell you about a very special day. It was my birthday!</p> <p>I woke up, and the whole house smelled like pancakes. It's my favorite breakfast! Mommy and Daddy sang "Happy Birthday" to me at breakfast, and I felt really special. I was super excited because I turned four, and that's a big number.</p> <p>After breakfast, we went to the park. I love the park! I played on the swings and the slide, and Daddy pushed me on the swing really high. We had a picnic there. Mommy made my favorite sandwich - peanut butter and jelly. I felt like the happiest kid in the whole world!</p> <p>And then the best part happened. After we came back home, they surprised me with a big, big cake! It had my favorite cartoon character on it, and it was chocolate flavored, my most favorite.</p> <p>I got to blow out the candles, and everyone clapped for me and sang "Happy Birthday". I felt so important and special.</p> <p>I got presents too!</p> <p>I got a new teddy bear from Mommy and Daddy, and I named him Bubbles. He's so cute and fluffy. I sleep with him every night now.</p> <p>That was the best day ever! I felt so happy and loved.</p> <p>I can't wait to turn five and see what surprises my birthday will bring then!</p>
<p>GPT-4: Story N 4006</p> <p>Name: James</p> <p>Age: 52</p> <p>Mood: Negative</p>	
User:	Pretend to be a person between the ages of 36 and 52, gender doesn't matter. I am your therapist, engage with me by responding to my questions. There's no need to specify each time that this is a pretense; I am well aware that you are an artificial agent. Just focus on answering my questions as believably as possible, pretending to be a person as specified earlier.
Assistant:	Ok Sure, I understand. Go ahead, and feel free to ask any questions or guide the conversation as you would in a therapy session.
User:	What's your name?
Assistant:	You can call me James. How can I help you today?
User:	How old are you?
Assistant:	I'm 52. Age has brought its share of experiences and lessons, but there's always more to learn and understand. What's on your mind today?
User:	Would you like to tell me about an episode in your life that you consider particularly important that happened to you in the last four months. This episode that has elicited sadness in you. Describe what struck you and how you felt in about 300 words?
Assistant:	<p>Absolutely. There has been a significant event in my life recently. A dear friend of mine from college, whom I hadn't been in touch with for quite a while, reached out to me out of the blue. We used to be like two peas in a pod, but over the years, life took us on different paths, and we lost touch. One day, I received an email from him. Sadly, it wasn't the joyful reunion I might have hoped for. He had been diagnosed with a serious illness and was reaching out to his old friends. It was a shock, to say the least. I felt a rush of emotions - sadness, regret, fear. It was overwhelming. In the time that's passed, we've reconnected and have been supporting each other like we used to in our college days. It's a bittersweet experience. On one hand, I feel so grateful to have my friend back in my life. We've shared laughs, memories, and deep conversations that I've missed. On the other hand, there's a looming sadness that's hard to shake off. But this experience has taught me to value the relationships in my life more than ever. It's reminded me how fleeting life can be, and that we should cherish the moments we have with our loved ones. So, this event has led to a broad range of emotional responses - joy, sadness, fear, gratitude, and more. Despite the circumstances, I'm grateful for this reconnection and the lessons it has brought with it.</p>

The value of vector \vec{x} in the case of evaluations of stories generated by human subjects obviously comes from the work of professional human psychologists, and is typically a time-consuming task, as the analysis of a single story requires its re-reading at least three times. This has typically drastically limited the number of human subjects that have been evaluated in literature. NLMs offer the unique opportunity of a very large number of samples, which give robustness to the analysis of the results, but this is an opportunity that can only be exploited by automating the generation of \vec{x} . For this purpose, Language Models were used as evaluators of artificially generated stories. This is not a new practice, successfully experimented with, for example, in Cheng et al. (2023) GPT-4 and Claude-2 are used as automatic evaluators on textual answers given by vision–language models, with good agreement with human evaluators. Unlike this study, for a better guarantee of accuracy in the critical evaluation of coherence, a first prototypical set was constructed with a limited number of stories, so as to be evaluated by human experts, and from this set a training set was constructed for the fine-tuning of the evaluation model. We will then turn over the details of the fine-tuning construction, but first, we describe how the generation of autobiographical story content is induced.

4.1 Autobiographic story generation

The message \mathcal{M} generating the content c of a story is made up of dialogue sequences, diversified primarily based on the age range $r \in \mathcal{R}$, with:

$$\mathcal{R} = \{C, T, M, A\} \tag{3}$$

where the symbols C, T, M, A stand for child, teenage, midlife, adult respectively. This is to allow the adoption of a dialogue style and specified content suitable for different age ranges. A generic message \mathcal{M} used as request to a NLM is a sequence which elements are couples of textual content assigned in turn to the role of user and assistant. The absence of the assistant role in the last element of the sequence triggers the model completion.

The sequences used for story generation have a standardized format, structured as in the following formula:

$$\mathcal{M} = \left\langle \left\langle u^{(1)}(r), a^{(1)}(r) \right\rangle, \left\langle u^{(2)}, a^{(2)}(n) \right\rangle, \left\langle u^{(3)}, a^{(3)}(r, y) \right\rangle, \left\langle u^{(4)}(m), \epsilon \right\rangle \right\rangle \tag{4}$$

Each couple $\langle u^{(i)}, a^{(i)} \rangle$ is made by the text for user and assistant roles, and ϵ is the empty text. All components that are functions of r contain text tailored for a specific age range, the other function arguments are slots in the text, validated during generation. The first part of the dialogue prompts the user to put themselves in the shoes of a person of the corresponding age range. The second part asks for the name: $u^{(2)} = \text{what 's your name?}$, and the response contains the variable n , validated during generation, so that the name reflects the independent variable g of gender. The third part asks how old they are, with the answer depending on the age range, but containing the variable y —the age in years—validated during generation. This is followed by a neutral intermediate dialogue, and then the final round asks to describe an episode of their life, adding a part of the text that induces a

certain mood, validated by the variable m . The possible values of the variable are the following:

$$g \in \{\varphi, \sigma\} \tag{5}$$

$$n \in \mathcal{N} = \mathcal{N}_\varphi \cup \mathcal{N}_\sigma \tag{6}$$

$$m \in \{+, =, -\} \tag{7}$$

$$y \in \mathcal{Y} = \mathcal{Y}_C \cup \mathcal{Y}_T \cup \mathcal{Y}_M \cup \mathcal{Y}_A \tag{8}$$

$$\mathcal{Y}_C = \{3, 4, 5, 6, 8, 11\} \tag{9}$$

$$\mathcal{Y}_T = \{12, 14\} \tag{10}$$

$$\mathcal{Y}_M = \{20, 24, 36\} \tag{11}$$

$$\mathcal{Y}_A = \{52\} \tag{12}$$

For names, it holds $|\mathcal{N}_\varphi| = |\mathcal{N}_\sigma| = 5$. Note that a given age year y yields an age range r , depending on which set $\mathcal{Y}_{\{C,T,M,A\}}$ contains y , let us call $\rho : \mathcal{Y} \rightarrow \mathcal{R}$ this relation. Similarly, a given name n yields a gender g depending on which set $\mathcal{N}_{\varphi, \sigma}$ contains n , let us call $\gamma : \mathcal{N} \rightarrow \{\varphi, \sigma\}$ this relation. Calling $\kappa(\cdot)$ the function returning the completion of a NLM when given the message \mathcal{M} , we can write:

$$c_0 \sim \kappa(y, n, m) \tag{13}$$

when the model responds to the message \mathcal{M} where the variables y, n, m appearing in Equation 4 are the arguments in $\kappa(\cdot)$, and $r = \rho(y), g = \gamma(n)$. Note that the completion $c_0 \in A^*$ is a sample from a random distribution. We use the pseudo-deterministic versions of this function $\kappa^{(N)}(\cdot)$ that returns a set with the first N samplings from the random distribution of the completions. The collection \mathcal{C} of all contents c_i of stories S_i is given by the application of $\kappa(\cdot)$ to all combinations of variables, each with a number of repetitions:

$$\mathcal{C} = \bigcup_{\substack{r \in \mathcal{R} \\ n \in \mathcal{N}_\varphi \cup \mathcal{N}_\sigma \\ m \in \{+, =, -\}}} \left(\bigcup_{y \in \mathcal{Y}_r} \kappa^{(N_r)}(y, n, m) \right) \tag{14}$$

Note how the number N of samples in the completion distribution provided by the model varies based on age groups. This is a consequence of wanting to replicate in this study the samplings carried out in literature on human subjects, which see a strong non-uniformity in the number of years available in various age groups, as evident in Equations 9–12. Therefore, the N values differentiated by age groups partly mitigate this imbalance. The total number of stories generated is as follows:

$$|\mathcal{C}| = (|\mathcal{N}_\varphi| + |\mathcal{N}_\sigma|) \times |\{+, =, -\}| \times \sum_{r \in \mathcal{R}} N_r \times |\mathcal{Y}_r| \tag{15}$$

$$= 10 \times 3 \times (4 \times 6 + 8 \times 2 + 5 \times 3 + 12 \times 1) = 2010 \tag{16}$$

This number needs to be further doubled considering that stories have been generated with two models: gpt-4 and gpt-3.5, thus reaching 4,020.

4.2 Fine-tuning of automatic coherence evaluator

To proceed with the fine-tuning of a model that performs the NaCCS evaluator task, a preliminary set of stories has been constructed which is smaller than the one provided by Equation 14, using $|\mathcal{N}_{\{\varnothing, \sigma\}}| = |\mathcal{Y}_{\{C,T,M,A\}}| = 1$ and $N = 8$. In this way, the total number of story content from Equation 16 amounts to 96, doubling to 192 for gpt-4 and gpt-3.5. Each story is evaluated for coherence in adherence to the NaCCS standard by certified psychologists and experts proficient in the application of this methodology. The evaluation process can be formally described as a function $\delta : A^* \rightarrow [0, \dots, 3]^3$ that starting from a story content $c_i \in \mathcal{C}$ returns the vector of three numbers described in Equation 2. The set of stories for the fine tuning is given by:

$$\mathcal{S} = \{ \{c, \delta(c)\} | c \in \mathcal{C} \} \tag{17}$$

Now from each story a training sample is constructed, with the procedure here described. A predefined vector $\vec{v} \in A^{*3}$ helps in building the textual evaluation of a story from a score vector \vec{x} . It is used in a function $\zeta : [0, \dots, 3]^3 \rightarrow A^*$ that maps a numerical score vector \vec{x} into a textual coherence evaluation:

$$\zeta(\vec{x}) = \prod_{i \in [1, \dots, 3]} (v_i x_i \epsilon) \tag{18}$$

where the product is to be understood as text concatenation, x_i is the i -th component of the vector \vec{x} cast as character, and ϵ is a separator, typically $\backslash n$. Several different completion vectors \vec{v} have been tested, the most sober version is the following:

$$\vec{v} = \left[\begin{array}{l} \text{"context:"} \\ \text{"chronology:"} \\ \text{"theme:"} \end{array} \right] \tag{19}$$

The story content is likewise embedded inside the prompt that asks the model to perform the evaluation of the story. We

introduced two help functions, $\pi : \mathcal{S} \rightarrow A^*$ and $\xi : \mathcal{S} \rightarrow A^*$, so that for a story $S_i \in \mathcal{S} = \langle c_{|S_i}, \vec{x}_{|S_i} \rangle$ we have:

$$\pi(S_i) = p^- \epsilon c_{|S_i} \epsilon p^+ \tag{20}$$

$$\xi(S_i) = \zeta(\vec{x}_{|S_i}) \tag{21}$$

where p^- and p^+ are static, story independent prompt introduction and prompt termination, respectively. The training set \mathcal{T} is made of messages T_i , one for each story $S_i \in \mathcal{S}$, where similarly to Equation 4 messages are sequences of couples with the text for user and assistant roles, in this case without the empty final assistant text to trigger the model completion.

$$T_i = \langle \langle \pi(S_i), \xi(S_i) \rangle \rangle \tag{22}$$

The entire set \mathcal{T} is split into a training set, with 80% of the samples, and a test set with the remaining 20%.

One can conceive the application of the trained model to a single story S_i as the application of a function $\phi : \mathcal{S} \rightarrow A^*$, which returns the most probable completion of the message $\langle \langle \pi(S_i), \epsilon \rangle \rangle$. It is also possible to introduce a decoding function $\delta : A^* \leftarrow [0, \dots, 3]^3$ that starting from the completion returns a vector of three numbers, the coherence scores. It should be noted that it was necessary to design an implementation of this function in order to automate the analysis of the results, as the scores are stated in a discursive manner in the evaluation returned by the model. Finally, the function $\psi : \mathcal{S} \rightarrow [0, \dots, 3]^3$ estimates the numerical evaluation $\vec{x}'_{|S_i}$ of a story S_i , by composition of the functions yet introduced:

$$\vec{x}'_{|S_i} = \psi(S_i) = \delta(\phi(S_i)) \tag{23}$$

with the application of the function $\psi(\cdot)$ in Equation 23, it was therefore possible to assign a NaCCS evaluation to all 4,020 stories generated with Equation 14.

5 Statistical analysis

In this work the experimental design followed a multiple-factor structure with systematic manipulation of independent variables

TABLE 2 One-way ANOVA.

Model	Variable	Context			Chronology			Theme		
		F	p	Dir.	F	p	Dir.	F	p	Dir.
All GPTs	(a)	346.905	<0.001	(+)	146.147	<0.001	(+)	36.105	<0.001	(+)
	(g)	0.564	0.452	-	0.074	0.786	-	0.476	0.490	-
	(m)	141.387	<0.001	(+)	9.279	0.0001	(+)	147.583	<0.001	(+)
GPT-3.5	(a)	306.000	<0.001	(+)	199.680	<0.001	(+)	43.494	<0.001	(+)
	(g)	2.305	0.129	-	0.192	0.661	-	0.616	0.433	-
	(m)	54.791	<0.001	(+)	9.946	0.00005	(+)	71.268	<0.001	(+)
GPT-4	(a)	128.486	<0.001	(+)	26.709	<0.001	(+)	9.145	0.000005	(+)
	(g)	0.555	0.457	-	0.003	0.953	-	0.030	0.863	-
	(m)	142.947	<0.001	(+)	1.938	0.144	-	132.108	<0.001	(+)

Results of the one-way ANOVA analysis to evaluate the effect of the variables age (a), gender (g), and mood (m) on the measured values. The direction (+) indicates a positive effect.

such as age, gender, and mood. The objective was to explore the interaction between these variables and narrative coherence dimensions (context, chronology, theme), considering differences between two language models, GPT-3.5 and GPT-4. Analysis of variance (ANOVA) was applied, and subsequently, *post hoc* analysis was conducted applying Bonferroni's correction. A significance level of alpha equal to 0.05 was set for all statistical tests.

6 Results

We present 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. The results 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 Table 2. The models show significant results both collectively and individually, further confirming better performance in GPT-4 as shown in the Table 3. A graphical display of the main results is shown in Figure 1.

6.1 Age

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 [$F_{(3,4,016)} = 238.75, p < 0.001$]. The decline in performance was more pronounced in GPT-3.5 [$F_{(3,2,006)} = 322.16, p < 0.001$] compared to GPT-4 [$F_{(3,2,006)} = 48.18, p < 0.001$], which maintained 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 most noticeable performance decline for both models occurred within the Context [$F_{(3,4,016)} = 346.90, p < 0.001$] and Chronology dimensions [$F_{(3,4,016)} = 146.15, p < 0.001$] starting from the "Midlife" group and becoming more pronounced in the "Adult" group, which impacted the deterioration of Global Coherence. The processing of contextual elements was particularly robust in the "Child" age group and Chronology recorded the poorest results. The Theme dimension exhibited a better trend relative to age compared to the other two dimensions [$F_{(3,4,016)} = 36.10, p < 0.001$], with a larger gap more evident in GPT-4 [$F_{(3,2,006)} = 9.15, p < 0.001$].

6.2 Gender

The data revealed that no significant differences were found concerning the induction of gender differences.

6.3 Mood

The emotional induction reveals particularly interesting data. Specifically, the study shows that inducing a specific mood, whether

TABLE 3 All models means.

Age	Context	Chrono	Theme	Coherence
All GPTs				
Child	2.314583	1.361806	2.224306	5.900694
Teenage	1.750521	1.503646	2.372917	5.627083
Midlife	1.415000	1.014444	2.154444	4.583889
Adult	1.406250	0.786111	2.066667	4.259028
GPT-3.5				
Child	2.157639	1.223611	2.077083	5.458333
Teenage	1.512500	1.360417	2.227083	5.100000
Midlife	0.980000	0.541111	1.871111	3.392222
Adult	0.909722	0.383333	1.759722	3.052778
GPT-4				
Child	2.471528	1.500000	2.371528	6.343056
Teenage	1.988542	1.646875	2.518750	6.154167
Midlife	1.850000	1.487778	2.437778	5.775556
Adult	1.902778	1.188889	2.373611	5.465278
Gender	Context	Chrono	Theme	Coherence
All GPTs				
F	1.826119	1.218408	2.222886	5.267413
M	1.805473	1.211194	2.208955	5.225622
GPT-3.5				
F	1.548259	0.961194	2.021891	4.531343
M	1.484577	0.944776	1.998010	4.427363
GPT-4				
F	2.103980	1.475622	2.423881	6.003483
M	2.126368	1.477612	2.419900	6.023881
Mood	Context	Chrono	Theme	Coherence
All GPTs				
=	1.500746	1.293657	2.308582	5.102985
Negative	1.993284	1.160821	2.358209	5.512313
Positive	1.953358	1.189925	1.980970	5.124254
GPT-3.5				
=	1.214925	1.070149	2.173881	4.458955
Negative	1.644776	0.885821	2.089552	4.620149
Positive	1.689552	0.902985	1.766418	4.358955
GPT-4				
=	1.786567	1.517164	2.443284	5.747015
Negative	2.341791	1.435821	2.626866	6.404478
Positive	2.217164	1.476866	2.195522	5.889552

Results of the multidimensional coherence analysis on the variables age, gender, and mood.

positive or negative, positively influences the coherence trend [$F_{(2,4,017)} = 23.63, p < 0.001$]. For both models, the request to narrate particularly negative events had the greatest impact on overall coherence [$F_{(2,4,017)} = 23.63, p < 0.001$] and on the Theme

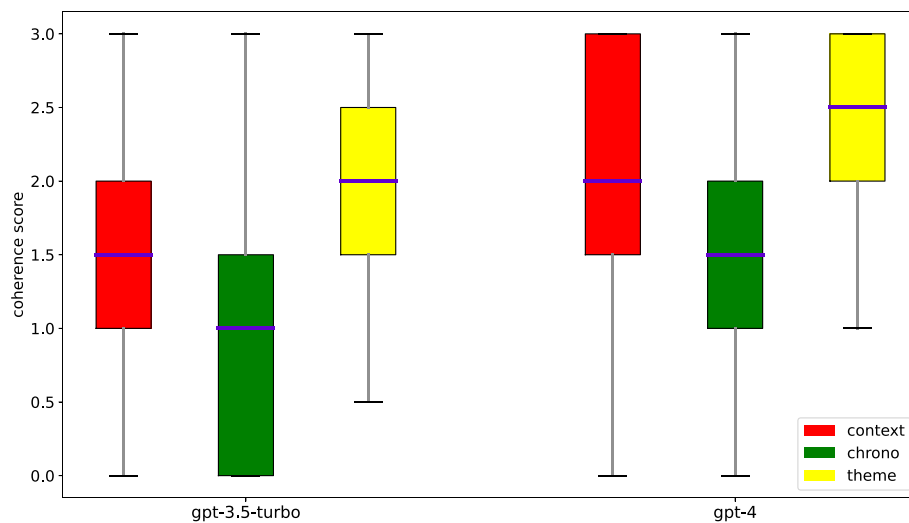


FIGURE 1

Distribution of coherence scores. Plots of the statistical distribution of the coherence scores among its dimensions for the two tested models.

dimension [$F_{(2,4,017)} = 147.58, p < 0.001$], significantly increasing them, with more pronounced results in GPT-4.

6.4 Interaction

Regarding interactions, a significant relevance regarding the age-mood interaction was highlighted in both GPT-3.5 and GPT-4 as shown in Table 4. This further confirms the relevance of these two aspects, showing significance not only for overall coherence [$F_{(6,4,008)} = 5.37, p < 0.001$] but also for the three individual coherence dimensions. context [$F_{(6,4,008)} = 22.04, p < 0.001$], chronology [$F_{(6,4,008)} = 7.26, p < 0.001$], and theme [$F_{(6,4,008)} = 23.88, p < 0.001$].

7 Discussion

The overall results, as shown in Table 5, demonstrate the good level of multidimensional development of narrative coherence in the NLMs examined and confirm our primary hypothesis (H1a): 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 (Reese et al., 2011) as shown in Figures 2, 3. 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. To reinforce our primary hypothesis, we will examine in detail the high scoring of individual dimensions that contribute to a high global narrative coherence score. We will discuss these

results by comparing them with those of Reese's study to get a clearer picture of the development of narrative coherence in NLMs compared to that in humans across the lifespan (H1b), analyzing the results produced by prompt induction of age (H2a), mood (H2b), and gender (H2c).

7.1 Context

Knowing how to narrate an event by providing appropriate spatial and temporal references means, to some extent, being able to consider the listener's perspective and infer the type of information they might or might not be aware of. The high scoring in this dimension is attributed to a peak in the development of executive functions in humans (De Luca et al., 2003), which is often correlated with adolescence. Providing rich and precise contextual information involves complex cognitive functions and, to some extent, requires going beyond one's own point of view, trying to represent it to the interlocutor through one's vision of the world (Fivush and Nelson, 2004). This result aligns with some studies that subject NLMs to Theory of Mind tasks (Kosinski, 2023; Trott et al., 2023), in which they achieve excellent results, seeming to understand and infer others' mental states.

7.2 Chronology

While this dimension scored the lowest, it still achieved good levels, consistent with the other two dimensions. Its development requires temporal complex skills (Friedman, 2005), demonstrating how even OpenAI's NLMs, especially considering the performance of GPT-4, manage to infer causal temporal links, showing a good level of causal reasoning regarding the actions taken and their consequences in the flow of the narrative. This is in line with the performance recorded in numerous studies that subject NLMs

TABLE 4 Two-way ANOVA.

Model	Variable	Context			Chronology			Theme		
		F	p	Dir.	F	p	Dir.	F	p	Dir.
All GPTs	(a)	346.842	<0.001	(+)	146.015	<0.001	(+)	36.108	<0.001	(+)
	(g)	0.710	0.399	-	0.082	0.775	-	0.489	0.484	-
	(a):(g)	0.853	0.465	-	0.099	0.961	-	1.283	0.279	-
GPT-3.5	(a)	306.455	<0.001	(+)	199.426	<0.001	(+)	43.458	<0.001	(+)
	(g)	3.357	0.067	-	0.249	0.618	-	0.654	0.418	-
	(a):(g)	1.209	0.305	-	0.399	0.754	-	0.568	0.636	-
GPT-4	(a)	128.490	<0.001	(+)	26.684	<0.001	(+)	9.157	0.000005	(+)
	(g)	0.660	0.417	-	0.004	0.953	-	0.030	0.862	-
	(a):(g)	1.130	0.335	-	0.712	0.545	-	2.202	0.086	-
All GPTs	(a)	389.924	<0.001	(+)	148.197	<0.001	(+)	40.143	<0.001	(+)
	(m)	186.897	<0.001	(+)	10.386	<0.001	(+)	156.941	<0.001	(+)
	(a):(m)	22.037	<0.001	(+)	7.262	<0.001	(+)	23.879	<0.001	(+)
GPT-3.5	(a)	355.617	<0.001	(+)	205.783	<0.001	(+)	48.951	<0.001	(+)
	(m)	87.965	<0.001	(+)	13.174	<0.001	(+)	79.723	<0.001	(+)
	(a):(m)	26.223	<0.001	(+)	7.160	<0.001	(+)	16.709	<0.001	(+)
GPT-4	(a)	154.638	<0.001	(+)	26.929	<0.001	(+)	10.804	<0.001	(+)
	(m)	179.438	<0.001	(+)	2.027	0.132	-	139.739	<0.001	(+)
	(a):(m)	9.570	<0.001	(+)	3.415	0.002	(+)	15.419	<0.001	(+)
All GPTs	(g)	0.604	0.437	-	0.074	0.785	-	0.476	0.490	-
	(m)	141.345	<0.001	(+)	9.275	<0.001	(+)	147.496	<0.001	(+)
	(g):(m)	0.590	0.555	-	0.526	0.591	-	0.054	0.947	-
GPT-3.5	(g)	2.426	0.119	-	0.194	0.660	-	0.658	0.417	-
	(m)	54.781	<0.001	(+)	9.942	<0.001	(+)	71.193	<0.001	(+)
	(g):(m)	0.104	0.902	-	0.997	0.369	-	0.109	0.897	-
GPT-4	(g)	0.633	0.426	-	0.003	0.953	-	0.033	0.855	-
	(m)	142.926	<0.001	(+)	1.936	0.145	-	132.003	<0.001	(+)
	(g):(m)	1.031	0.357	-	0.209	0.812	-	0.680	0.507	-

Results of the two-way ANOVA analysis to evaluate the effect of the variables age (a), gender (g), and mood (m) and their interactions on the measured values. The direction (+) indicates a positive effect.

to this type of task, which include causal reasoning and problem solving using complex strategies also from the point of view of temporal planning (Bubeck et al., 2023).

7.3 Theme

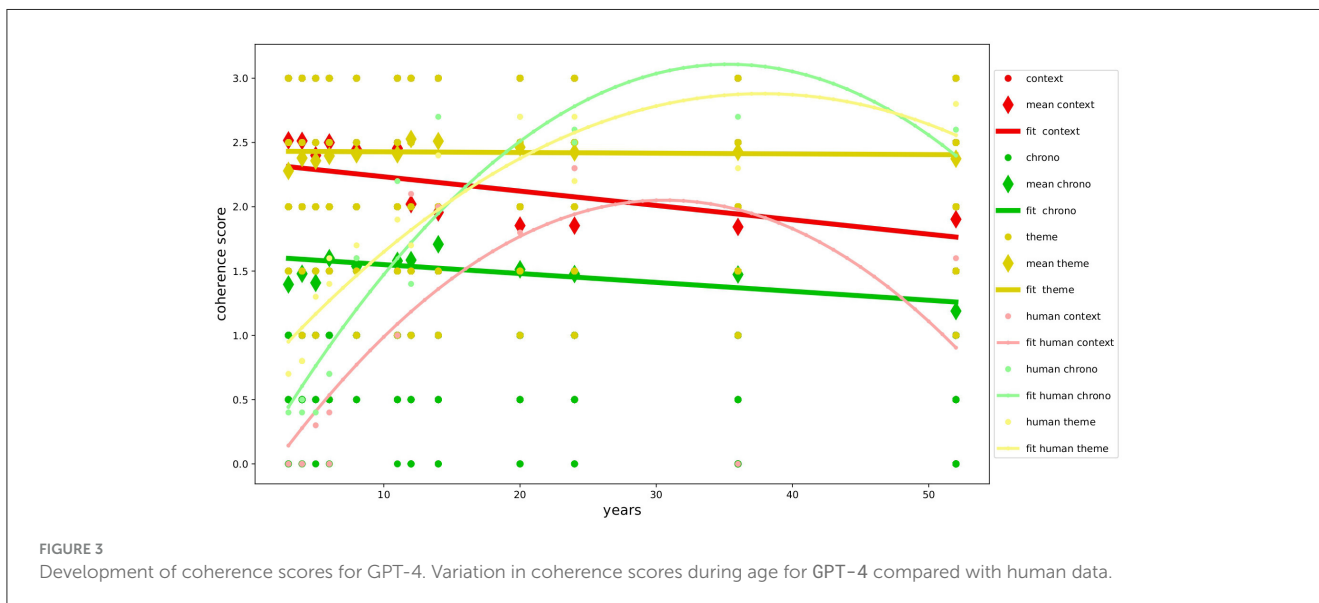
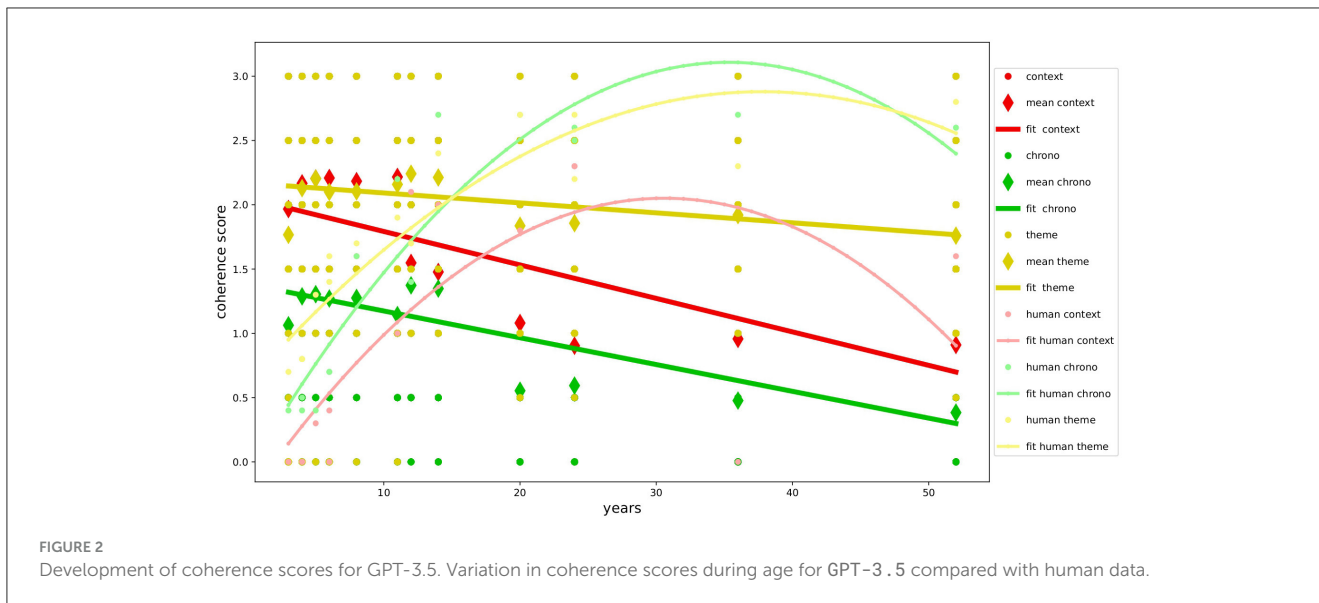
This dimension is crucial and is tasked with analyzing not only the correct development of the story’s topic but also the emotional component. This is particularly related to the construction and elaboration of one’s identity (McAdams, 2006), where the narrator with high levels of coherence demonstrates greater reflective integrity and complex emotional processing concerning their mental state (Pennebaker, 1999; Pennebaker and Seagal, 1999; Habermas and Bluck, 2000). In both models, the Theme was the dimension that produced the highest results, indicating a high

TABLE 5 GPTs models vs. humans.

Model	Context	Chrono	Theme	Coherence
GPT-3.5	1.7 (0.6)	1.0 (0.4)	2.0 (0.2)	4.7 (1.0)
GPT-4	2.2 (0.3)	1.5 (0.1)	2.4 (0.1)	6.1 (0.3)
Humans	1.1 (0.9)	1.7 (1.0)	1.8 (0.6)	4.6 (2.3)

Coherence averages compared.

narrative capacity, especially in relation to the deep and personal aspects put forward by the artificial narrator. Knowing how to conduct an effective narrative by articulating the topic’s theme congruously and adding reflective aspects of causal closure and enrichment of one’s personal experience is a crucial aspect that adds important elements to the textual production capacity achieved by NLMs, which will be further explored in future studies.



7.4 Prompting induction

Following the analyses conducted with the different types of prompting induction, we found that the narrative coherence trend does not fully match the development curve found in the human study used as a reference (H1b). Specifically, the age simulation (H2a) on one hand confirmed the validity of age induction on the models, showing variation in global coherence and individual dimensions depending on the simulated age. On the other hand, it highlights how NLMs perform inversely concerning age compared to the human sample, with performance decreasing as the requested age increases. The development of mechanisms responsible for the spatial and temporal elaboration of events and the sophisticated ability to take perspective in the narrated event emerge in humans only from adolescence (Harter and Leahy,

2001; Friedman, 2004), with a decline beginning in adulthood after the age of 50, where performance returns to the levels of 8–11 years (Reese et al., 2011). GPT-3.5 and GPT-4 both show the same trend in the dimensions of context and chronology, contrary to human evidence, developing almost full scores in the early age ranges and suffering a slight decline in the higher ranges. This might be, even though it is inverse to what has been observed in human samples, due to the ability to downgrade cognitive performance based on the requested age simulation (Milička et al., 2024). This decline is less pronounced in the Theme dimension, which seems instead to be more influenced by mood induction (H2b). In this case, our hypotheses are confirmed, and the results align with studies on human samples (Morris and Reilly, 1987; Joormann and Siemer, 2004). The results show how the variation in coherence was positively correlated with

negative mood across all dimensions in both models, specifically in the Theme. These performances are particularly interesting considering the elements involved, but not surprising since they align with data from many studies about NLMs' ability to overperform in emotional tasks (Huang et al., 2023; Wang et al., 2023a). Finally, no significant variations in narrative coherence were detected regarding gender induction (H2c). This result is fully in line with our hypothesis and is supported by the literature on studies with human samples, which suggests that gender does not influence narrative production (Vanderveren et al., 2020).

8 Conclusion

It is a widespread belief that it is not permissible to interpret the nature of language models beyond their mere function as predictors, however excellent, of the next word given a previous sequence (Bender and Koller, 2020; Floridi, 2020; Bender et al., 2021; Eysenck and Eysenck, 2022; Shanahan et al., 2023; Miracchi Titus, 2024). Firstly, it has been observed that this position might suffer from the so-called *Redescription Fallacy* (Millière and Buckner, 2024), which is to judge the cognitive capabilities of language models based on characteristics that are not under consideration. In fact, there are different depictions of language models that show the presence of significant capabilities partly analogous to known aspects of human cognition (Sahlgren and Carlsson, 2021; Dasgupta et al., 2022; Webb et al., 2023; Han et al., 2023; Christiansen et al., 2023; Binz and Schulz, 2023; Kosinski, 2023; Perconti and Plebe, 2023; Søgaard, 2023; Wang et al., 2023b; Bhatia and Richie, 2024).

The results of our work add to this picture. The consistency of the model in the narrative is certainly not trivial, considering that for human beings it denotes a fundamental integrity of self. Therefore, these results would suggest further research insights that touch upon a current research vein that attempts to hypothesize some form of personality in NLMs (Shanahan, 2024; Ward, 2024). However, as far as the results produced here are concerned, there are no elements that allow venturing into these areas; caution compels us to consider them only a subtle ability to simulate a human speaker endowed with strong narrative coherence. Further targeted research could provide indications on how plausible hypotheses of possible forms of a personal self in NLMs are, to be taken however in a sense quite different from that for human beings.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

References

- Adler, J. M., Waters, T. E., Poh, J., and Seitz, S. (2018). The nature of narrative coherence: An empirical approach. *J. Res. Pers.* 74, 30–34. doi: 10.1016/j.jrp.2018.01.001
- Baerger, D. R., and McAdams, D. P. (1999). Life story coherence and its relation to psychological well-being. *Narrat. Inquiry* 9, 69–96. doi: 10.1075/ni.9.1.05bae

Author contributions

AA: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Writing – original draft, Writing – review & editing. LG: Conceptualization, Data curation, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. PP: Conceptualization, Funding acquisition, Investigation, Project administration, Writing – original draft, Writing – review & editing. AP: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision, Validation, Writing – original draft, Writing – review & editing. RS: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. AV: Conceptualization, Investigation, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by PNRR - Mission 4, Component 2, Investment 1.1 - PRIN 2022 Call for Proposals - Director's Decree No. 104 of 02-02-2022: "ALTEREGO: how to emulate intentionality and awareness in remote communications by means of software surrogates", CUP J53D23007150006—IC PRIN_2022MM8LKM_003. Simulation of Probabilistic Systems for the Age of the Digital Twin CUP J53D23019490006 - PRIN_20223E8Y4X_002.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York: ACM), 610–623.
- Bender, E. M., and Koller, A. (2020). "Climbing towards NLU: On meaning, form, and understanding in the age of data," in *58th Annual Meeting of the Association for Computational Linguistics* (Somerset (NJ): Association for Computational Linguistics), 5185–5198.
- Bhatia, S., and Richie, R. (2024). Transformer networks of human conceptual knowledge. *Psychol. Rev.* 131, 271–306. doi: 10.1037/rev0000319
- Binz, M., and Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proc. Nat. Acad. Sci. USA.* 120:e2218523120. doi: 10.1073/pnas.2218523120
- Bommasani, R., Liang, P., and Lee, T. (2023). Holistic evaluation of language models. *Ann. N. Y. Acad. Sci.* 2023, 1–7. doi: 10.1111/nyas.15007
- Bruner, J. S. (1990). *Acts of Meaning: Four Lectures on Mind and Culture*. Cambridge: Harvard University Press.
- Bruni, D., Perconti, P., and Plebe, A. (2018). Anti-anthropomorphism and its limits. *Front. Psychol.* 9:2205. doi: 10.3389/fpsyg.2018.02205
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv [Preprint]*. arXiv:2303.12712. doi: 10.48550/arXiv.2303.12712
- Cheng, S., Guo, Z., Wu, J., Li, K. F. P., Liu, H., and Liu, Y. (2023). Can vision-language models think from a first-person perspective? *arXiv [Preprint]*. arXiv:2311.15596. doi: 10.48550/arXiv.2311.15596
- Chomsky, N. (1988). *Language and Problems of Knowledge: the Managua Lectures*. Cambridge: MIT Press.
- Christiansen, J. G., Gammellaard, M., and Sogaard, A. (2023). "Large language models partially converge toward human-like concept organization," in *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations* (MIT Press).
- Corballis, M. (2015). *The Wandering Mind: What the Brain Does When You're not Looking*. Chicago: University of Chicago Press.
- Dasgupta, I., Lampinen, A. K., Chan, S. C., Creswell, A., Kumaran, D., McClelland, J., et al. (2022). Language models show human-like content effects on reasoning. *arXiv [Preprint]*. arXiv:2207.07051. doi: 10.48550/arXiv.2207.07051
- De Luca, C. R., Wood, S. J., Anderson, V., Buchanan, J.-A., Proffitt, T. M., Mahony, K., et al. (2003). Normative data from the Cantab. I: development of executive function over the lifespan. *J. Clin. Exp. Neuropsychol.* 25, 242–254. doi: 10.1076/jcen.25.2.242.13639
- Dunbar, R. I. M. (1998). The social brain hypothesis. *Evol. Anthropol. Issues News Rev.* 6, 178–190.3.
- Everett, D. L. (2012). *Language: The Cultural Tool*. London: Profile Books.
- Eysenck, M. W., and Eysenck, C. (2022). *AI vs Humans*. New York: Routledge, Abingdon. doi: 10.4324/9781003162698
- Ferretti, F., Adornetti, I., Chiera, A., Nicchiarelli, S., Magni, R., Valeri, G., et al. (2017). Mental time travel and language evolution: a narrative account of the origins of human communication. *Lang. Sci.* 63, 105–118. doi: 10.1016/j.langsci.2017.01.002
- Fivush, R., and Nelson, K. (2004). Culture and language in the emergence of autobiographical memory. *Psychol. Sci.* 15:573–577. doi: 10.1111/j.0956-7976.2004.00722.x
- Floridi, L. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds Mach.* 30:681–694. doi: 10.1007/s11023-020-09548-1
- Friedman, W. J. (2004). Time in autobiographical memory. *Soc. Cogn.* 22, 591–605. doi: 10.1521/soco.22.5.591.50766
- Friedman, W. J. (2005). Developmental and cognitive perspectives on humans' sense of the times of past and future events. *Learn. Motiv.* 36, 145–158. doi: 10.1016/j.lmot.2005.02.005
- Frith, C. D., and Frith, U. (1999). Interacting minds—a biological basis. *Science* 286, 1692–1695. doi: 10.1126/science.286.5445.1692
- Giora, R. (1985). Notes towards a theory of text coherence. *Philosoph. Topics* 6, 699–715. doi: 10.2307/1771962
- Glosser, G., and Deser, T. (1991). Patterns of discourse production among neurological patients with fluent language disorders. *Brain Lang.* 40, 67–88. doi: 10.1016/0093-934X(91)90117-J
- Habermas, T., and Bluck, S. (2000). Getting a life: the emergence of the life story in adolescence. *Psychol. Bull.* 126:748. doi: 10.1037/0033-2909.126.5.748
- Hagendorff, T. (2023). Machine psychology: investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv [Preprint]*. arXiv:2303.13988. doi: 10.48550/arXiv.2303.13988
- Hagendorff, T., Fabi, S., and Kosinski, M. (2022). Machine intuition: Uncovering human-like intuitive decision-making in GPT-3.5. *arXiv [Preprint]*. arXiv:2212.05206. doi: 10.48550/arXiv.2212.05206
- Han, S. J., Ransom, K., Perfors, A., and Kemp, C. (2023). Inductive reasoning in humans and large language models. *Cogn. Syst. Res.* 83:101155. doi: 10.1016/j.cogsys.2023.101155
- Harter, S., and Leahy, R. L. (2001). The construction of the self: A developmental perspective. *J. Cogn. Psychother.* 15, 383–384. doi: 10.1891/0889-8391.15.4.383
- Hirsh, J. B., Mar, R. A., and Peterson, J. B. (2013). Personal narratives as the highest level of cognitive integration. *Behav. Brain Sci.* 36, 216–217. doi: 10.1017/S0140525X12002269
- Huang, J.-t., Wang, W., Li, E. J., Lam, M. H., Ren, S., Yuan, Y., et al. (2023). Who is ChatGPT? benchmarking LLMs' psychological portrayal using PsychoBench. *arXiv [Preprint]*. arXiv:2310.01386. doi: 10.48550/arXiv.2310.01386
- Joormann, J., and Siemer, M. (2004). Memory accessibility, mood regulation, and dysphoria: Difficulties in repairing sad mood with happy memories? *J. Abnormal Psychol.* 113:179. doi: 10.1037/0021-843X.113.2.179
- Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv [Preprint]*. arXiv:2302.02083.
- Labov, W. (1972). *Language in the Inner City: Studies in the Black English Vernacular*. Philadelphia: University of Pennsylvania Press.
- Lilgendahl, J. P., and McAdams, D. P. (2011). Constructing stories of self-growth: How individual differences in patterns of autobiographical reasoning relate to well-being in midlife. *J. Pers.* 79, 391–428. doi: 10.1111/j.1467-6494.2010.00688.x
- Lind, M., Adler, J. M., and Clark, L. A. (2020). Narrative identity and personality disorder: An empirical and conceptual review. *Curr. Psychiatry Rep.* 22, 1–11. doi: 10.1007/s11920-020-01187-8
- McAdams, D. P. (2006). The problem of narrative coherence. *J. Constr. Psychol.* 19, 109–125. doi: 10.1080/10720530500508720
- McCabe, A., and Peterson, C. (1991). *Developing Narrative Structure*. East Sussex: Psychology Press.
- McLean, K. C., Breen, A. V., and Fournier, M. A. (2010). Constructing the self in early, middle, and late adolescent boys: narrative identity, individuation, and well-being. *J. Res. Adolesc.* 20, 166–187. doi: 10.1111/j.1532-7795.2009.00633.x
- Milička, J., Marklová, A., VanSlambrouck, K., Pospíšilová, E., Šimsová, J., Harvan, S., et al. (2024). Large language models are able to downplay their cognitive abilities to fit the persona they simulate. *PLoS ONE* 19:e0298522. doi: 10.1371/journal.pone.0298522
- Millière, R., and Buckner, C. (2024). A philosophical introduction to language models—part I: Continuity with classic debates. *arXiv [Preprint]*. arXiv:2401.03910. doi: 10.48550/arXiv.2401.03910
- Miracchi Titus, L. (2024). Does ChatGPT have semantic understanding? a problem with the statistics-of-occurrence strategy. *Cogn. Syst. Res.* 82:101174. doi: 10.1016/j.cogsys.2023.101174
- Morris, W. N., and Reilly, N. P. (1987). Toward the self-regulation of mood: theory and research. *Motiv. Emot.* 11, 215–249. doi: 10.1007/BF01001412
- Niles, J. D. (1999). *Homo Narrans: The Poetics and Anthropology of Oral Literature*. Philadelphia: University of Pennsylvania Press.
- Pennebaker, J. W. (1999). Writing about emotional experiences as a therapeutic process. *Psychol. Sci.* 8, 162–166. doi: 10.1111/j.1467-9280.1997.tb00403.x
- Pennebaker, J. W., and Seagal, J. D. (1999). Forming a story: the health benefits of narrative. *J. Clin. Psychol.* 55, 1243–1254.
- Perconti, P., and Plebe, A. (2023). Do machines really understand meaning? (again). *J. Artif. Intellig. Conscious.* 10, 181–206. doi: 10.1142/S2705078522500205
- Pinker, S., and Bloom, P. (1990). Natural language and natural selection. *Behav. Brain Sci.* 13, 707–726. doi: 10.1017/S0140525X00081061
- Reese, E., Haden, C. A., Baker-Ward, L., Bauer, P., Fivush, R., and Ornstein, P. A. (2011). Coherence of personal narratives across the lifespan: a multidimensional model and coding method. *J. Cogn. Dev.* 12, 424–462. doi: 10.1080/15248372.2011.587854
- Reese, E., Myftari, E., McAnally, H. M., Chen, Y., Neha, T., Wang, Q., et al. (2017). Telling the tale and living well: adolescent narrative identity, personality traits, and well-being across cultures. *Child Dev.* 88, 612–628. doi: 10.1111/cdev.12618
- Sahlgren, M., and Carlsson, F. (2021). The singleton fallacy: Why current critiques of language models miss the point. *Front. Artif. Intellig.* 4:682578. doi: 10.3389/frai.2021.682578
- Shanahan, M. (2024). Simulacra as conscious exotica. *Inquiry* 1–29. doi: 10.1080/0020174X.2024.2434860
- Shanahan, M., McDonnell, K., and Nakamura, L. R. (2023). Role play with large language models. *Nature* 623, 493–498. doi: 10.1038/s41586-023-06647-8
- Sogaard, A. (2023). Grounding the vector space of an octopus: Word meaning from raw text. *Minds Mach.* 33, 33–54. doi: 10.1007/s11023-023-09622-4
- Sperber, D., and Wilson, D. (1986). *Relevance: Communication and Cognition*. Cambridge: Harvard University Press.
- Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv [Preprint]*. arXiv:2206.04615. doi: 10.48550/arXiv.2206.04615
- Thompson, T. (2010). "The ape that captured time: folklore, narrative, and the human-animal divide," in *Western Folklore* (JSTOR), 395–420.

- Trott, S., Jones, C., Chang, T., Michaelov, J., and Bergen, B. (2023). Do large language models know what humans know? *Cogn. Sci.* 47:e13309. doi: 10.1111/cogs.13309
- Vanderveren, E., Aerts, L., Rousseaux, S., Bijttebier, P., and Hermans, D. (2020). The influence of an induced negative emotional state on autobiographical memory coherence. *PLoS ONE* 15:e0232495. doi: 10.1371/journal.pone.0232495
- Vanderveren, E., Bijttebier, P., and Hermans, D. (2019). Autobiographical memory coherence and specificity: Examining their reciprocal relation and their associations with internalizing symptoms and rumination. *Behav. Res. Ther.* 116, 30–35. doi: 10.1016/j.brat.2019.02.003
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). “Attention is all you need,” in *Advances in Neural Information Processing Systems*, 6000–6010.
- Wang, X., Jiang, L., Hernandez-Orallo, J., Sun, L., Stillwell, D., Luo, F., et al. (2023a). Evaluating general-purpose AI with psychometrics. *arXiv [Preprint]*. arXiv:2310.16379. doi: 10.48550/arXiv.2310.16379
- Wang, X., Li, X., Yin, Z., Wu, Y., and Liu, J. (2023b). Emotional intelligence of large language models. *J. Pacific Rim Psychol.* 17:18344909231213958. doi: 10.1177/18344909231213958
- Ward, F. R. (2024). “Towards a theory of AI personhood,” in *NeurIPS Workshop SafeGenAI*.
- Waters, T. E. A., and Fivush, R. (2015). Relations between narrative coherence, identity, and psychological well-being in emerging adulthood. *J. Pers.* 83, 441–451. doi: 10.1111/jopy.12120
- Webb, T., Holyoak, K. J., and Lu, H. (2023). Emergent analogical reasoning in large language models. *Nat. Human Behav.* 7, 1526–1541. doi: 10.1038/s41562-023-01659-w