



Transforming agency: On the mode of existence of large language models

Xabier E. Barandiaran¹ · Lola S. Almendros²

Accepted: 22 July 2025
© The Author(s) 2025

Abstract

This paper investigates the ontological characterization of Large Language Models (LLMs) like ChatGPT. Between inflationary and deflationary accounts, we pay special attention to their status as agents. This requires explaining in detail the architecture, processing, and training procedures that enable LLMs to display their capacities, and the extensions used to turn LLMs into agent-like systems. After a systematic analysis we conclude that a LLM fails to meet necessary and sufficient conditions for autonomous agency in the light of embodied theories of mind: the *individuality* condition (it is not the product of its own activity, it is not even directly affected by it), the *normativity* condition (it does not generate its own norms or goals), and, partially the *interactional asymmetry* condition (it is not the origin and sustained source of its interaction with the environment). If not agents, then... What are LLMs? We argue that ChatGPT should be characterized as an interlocutor or linguistic automaton, a *library-that-talks*, devoid of (autonomous) agency, but capable to engage performatively on non-purposeful yet purpose-structured and purpose-bounded tasks. When interacting with humans, a “ghostly” component of the human-machine interaction makes it possible to enact genuine conversational experiences with LLMs. Despite their lack of sensorimotor and biological embodiment, LLMs textual embodiment (the training *corpus*), digital extended interface embodiments, and resource-hungry computational embodiment, significantly transform existing forms of machine automatism and human agency. Beyond assisted and extended agency, the LLM-human coupling can produce *midtended* forms of agency, closer to the production of intentional agency than to the extended instrumentality of any previous technologies.

Keywords Transformers · Large language models · Agency · Autonomy · Interlocutor automata · Human machine interaction

Extended author information available on the last page of the article

1 Introduction

The recent emergence of Large Language Models (LLMs hereafter) (see Brown et al., 2020) with their wide availability¹ and their human-like generative capabilities are (re)opening the question around the ontological status of Artificial Intelligence. Are these systems genuinely intelligent? Do they possess mindful capacities? The responses are often polarized (Mitchell & Krakauer, 2023). Inflationary views (fuelled by research enthusiasm and commercial interest alike) tend to amplify AI properties, assimilating or approximating them to the human (and the superhuman). Deflationary views (typically trying to mitigate the harms of inflationary marketing), tend to downplay capacity attributions, and bring AI systems closer to dumb mathematical or mechanical devices. *Deflationary* categorizations typically revolve around treating LLMs as statistical processors, “stochastic parrots” (Bender et al., 2021), a “blurry JPEG of the web” (Chiang, 2023), “lumbering statistical engine for pattern matching” (Chomsky et al., 2023), “illusory thinking” (Shojaee et al., 2025) or simply “bullshit” (Hicks et al., 2024)². The most *inflationary* characterizations range from considering LLMs as “human-brain equivalents” (Ge et al., 2023), “genuine authors” and “accountable” entities (Miller, 2023), up to a “fully sentient person” (Lemoine, 2022). Somewhere in the middle stand more technical characterizations like “artificial reasoners” (Wei et al., 2023), “learners” (Brown et al., 2020), “general pattern machines” (Mirchandani et al., 2023), “sparky artificial general intelligence” (Bubeck et al., 2023) or simply “language models” with the slippery temptation to be turned into “world models” (K. Li et al., 2023).

An increasing danger of some deflationary views of LLMs is that, from their point of view, most of the risks can be attributed to the influence of misguided inflationary conceptions. It is often assumed that these can be mitigated if inflationary views are conclusively shown to be wrong: “Behind the smog of the hype and the marketing” the argument goes “there is no genuine intelligence or understanding behind LLMs, they are simple statistical processors, the only problem (besides their energy consumption and biases) is that other humans take them at face value”. Moreover, the argument continues, “if we treat them as the stupid machines they are”, the conclusion follows, “even the issues of bias and energy should fade away”. Or, as Chomsky et al. conclude, “Given the amorality, faux science and linguistic incompetence of these systems, we can only laugh or cry at their popularity” (2023). The real capacity of LLMs is thus left disregarded by such deflationary views, both as a potential risk to society and as a genuine source of positive sociotechnical transformation that needs to be more deeply

¹ Online services like Gemini, Claude, or, more prominently, ChatGPT, and the free/open source alternatives that can also run locally, like DeepSeek, Qwen, LLaMa, Mixtral, etc. are making these technologies massively available, not only as direct conversational bots but also, and importantly, as integrated assistants or agency boosting systems into different applications.

² Although all the cited deflationary accounts of what LLMs do provide critical arguments to defend their position, the term “bullshit,” as used by Hicks et al., is more technical and better defended than it might appear at first sight. Hicks et al. adopt Frankfurt’s (2005) explanation of “bullshit talk” as discourse that disregards truth without directly aiming to mislead the receiver or intentionally conceal it, as lying does. They show how this analysis applies to LLM outputs as well.

thought out. If we do not fine-tune our conceptualization of what LLMs are, we will not be able to properly analyze, evaluate, stir, and communicate their impact.

There are many ways of assessing this “real” power and its impact. Some are historic and socio-economic (Pasquinelli, 2023), or socio-ecological (Crawford, 2021), or even existential (Bostrom, 2017; Christian, 2021; Russell, 2019). But little attention has been put on critically analyzing LLMs from the point of view of “agency”. Despite the widespread academic consensus on the lack of conscious or sentient capabilities of LLMs, their status as agents is often uncritically assumed or proclaimed both in philosophy (Floridi, 2023) and the industry (OpenAI, 2025b). And there are two good reasons to strengthen accuracy on agentic attributions to LLMs: (a) mindful capacities did not arise in nature as a result of chess playing but of the evolution of agency (Barandiaran, 2008; Sterelny, 2001; Tomasello, 2022) and thus granting LLMs the status of agents means assuming them on a path towards such properties, and (b) achieving autonomous agency is one of the next big things in AI (Wang et al., 2023) with widespread support from Google, Apple, and Microsoft (including OpenAI) recently marketing their AI products as “agents” (Holmes, 2024; O’Donnell, 2024).; a point that became particularly relevant since the recent launch of ChatGPT Agent (OpenAI, 2025b).

In the next section, we contextualize the problem of the ontological status of LLM in terms of its current capabilities and limitations as expressed on different benchmarks. We show how transformer technologies are breaking down the solid distinctions between the human and the engineered. The *mode of existence* of these technologies is, however, not limited to their performance. In order to clarify it, we follow a threefold approach. First, in Sect. 3, we ground our analysis in the technical reality of these systems by focusing on the internal workings of ChatGPT-like LLMs: their architecture, processing flow, training procedures, and the extensions used to create agent-like behaviors. A clear technical understanding is the necessary foundation for any meaningful ontological assessment. Second, with these technical details at hand, in Sect. 4, we systematically evaluate their status as agents from the perspective of contemporary embodied and enactive theories of mind. We conclude that they fail to meet the core conditions for autonomous agency. Third, in Sect. 5, we argue that this negative conclusion is insufficient; it compels us to address the crucial question: If not agents, then what are they? Based on the technical specifications of Sect. 3, and building on embodied and enactive approaches to cognition, we characterize them as *interlocutor automata*, capable of bringing digital textual bodies to conversational life in interaction with humans, with the capacity to deeply transform human agency. We finally conclude and discuss the implications of our approach.

2 When computers can

Benchmarks, particularly when out of reach for the available technology, have often helped to reach agreement on the (lack of) capacities of AI systems. Now, many defend that the Turing Test is outdated (Bayne & Williams, 2023; Biever, 2023; Srivastava et al., 2023; Tikhonov & Yamshchikov, 2023). Generic conversations with

LLMs are indistinguishable from those we can enjoy with other humans (Jones & Bergen, 2024). More systematic variations of the Turing Test, directed at capturing the capacity of AI to display common sense, like the Winograd schema challenge (Levesque et al., 2012), have been declared obsolete (Kocijan et al., 2023). More sophisticated common sense reasoning tests like Winogrande (Sakaguchi et al., 2019) and HellaSwag (Zellers et al., 2019), specifically designed to be particularly hard for LLMs, have also been passed (Gemini Team et al., 2023; Jiang et al., 2023; OpenAI, 2023). Moreover, by 2023, LLMs, like GPT-4 already exhibited, according to their creators, “human-level performance” in a wide variety of professional and academic exams (OpenAI, 2023, p. 6), or, like Google’s Gemini, improving GPT-4 in many benchmarks (Gemini Team et al., 2023), could outperform humans on multitask language understanding tests (Hendrycks et al., 2021).

Current proprietary frontier systems—GPT-4o (OpenAI, 2024a), the GPT-o3 and 4o-mini (OpenAI, 2025a), Claude 4 family (Anthropic, 2025), and Gemini 2.5 Pro (Google DeepMind, 2025)—now reach or surpass human-level accuracy on many reasoning and commonsense suites. Open-source counterparts have rapidly closed the gap: KIMI K2 (Moonshot, 2025/2025), Llama 4 (Meta AI, 2025), Mixtral of Experts and its successor Magistral (Jiang et al., 2024a, b; Mistral-AI et al., 2025), DeepSeek-R1 (DeepSeek-AI et al., 2025), or Qwen 2.5 (Qwen et al., 2025) all report benchmark scores within a few percentage points of their closed-weight peers, while offering transparent and reproducible alternatives for academic research.

In concrete knowledge domains (like medicine) frontier models since 2023 often outperform average specialists at specific tasks (Guo et al., 2023; Singhal et al., 2025; Van Veen et al., 2024) and, perhaps not so surprisingly, can imitate philosophers with hardly distinguishable snippets (Schwitzgebel et al., 2024). More recently, methodological innovations on LLM’s self-guided “reasoning” (Lightman et al., 2023; Madaan et al., 2023a; Wei et al., 2023; Yao et al., 2023), and the resulting increase of inference-time compute, have pushed state-of-the-art scores by even higher margins (DeepSeek-AI et al., 2025; OpenAI, 2024b). A significant boost in benchmarking has also resulted from agentic extensions and tool use of LLMs (OpenAI, 2025b).

This rapid progress has accelerated the obsolescence of existing benchmarks, forcing the latest 1 M dollar prized Artificial General Intelligence benchmark (ARC-AGI) to be updated with a more challenging successor ARC-AGI2 in March 2025 (Chollet et al., 2025), and an even more challenging and agent focused ARC-AGI3 in July 2025³.

However, aggregate leaderboard numbers can hide “pockets of unpredictable failure” and thus overstate real-world reliability (Burnell et al., 2023). There are arguable limitations of current models, particularly in relation to some abstract reasoning capacities like compositionality (Dziri et al., 2023), multistep reasoning (Sprague et al., 2023), or complex planning (Valmeekam et al., 2023); whose mastery, by the way, is often rare among humans⁴. Large-scale studies show that scaling and instruc-

³ <https://three.arcprize.org>.

⁴ Most recently, Apple’s *The Illusion of Thinking* study (Shojaee et al., 2025) attracted extraordinary media attention in an increasingly polarized context. Subsequent analyses, however, have contested the

tion fine-tuning might not *fully* eliminate these weaknesses and could even make errors harder to predict (Zhou et al., 2024). Moreover, it is still possible, although increasingly harder, to find tasks in which humans score high, without specific expertise, yet remain hard for LLMs, as shown by the ARC-AGI-2 and 3 benchmarks (Chollet et al., 2025). And yet, what is certain is that we are facing the development of complex technologies that perform operations whose results are very similar to those requiring high levels of intelligence in humans. This circumstance translates into a growing undifferentiation between the human and the engineered. Conceptual divides that were once sharp and fixed are starting to melt and move. These advances force us to re-organize conceptual and ontological commitments regarding minds, machines, and agency.

Ever since the very conception of modern computers as universal Turing machines, the possibility of instantiating human, or super-human, level intelligence was at stake (Turing, 1950). At some point, this race between mind and machines settled down. Machines could (out)perform humans provided that the domain of interaction was rule-based, constrained, and limited, so that humans could program machines specifying the computational procedures required to carry out the task. Machines, however, were left with genuine mindfulness out of reach. Real-world tasks such as open conversations, a trip to the grocery, creative writing or subtle comforting humor were only within reach for us. The (human) mind could not be reduced to rule-following or explicit reasoning capacities but emerged, instead, out of sub-rational skillful embodied interactions that could not be implemented into machines (Dreyfus, 1992). Elephants, after all, don't play chess, but their mental life is rich and complex, nothing like computers would ever be able to accomplish (Brooks, 1990). On the other hand, computers could play chess but not pass the Turing Test, they could imitate and outperform humans in specific rule-based scenarios but not on the open field of language games and skilled conversations fuelled by a *common-sense*: an embodied sub-symbolic mesh that was claimed to resist rational, explicit, operationalization (Johnson & Lakoff, 2002; Varela et al., 1991).

Transformers have come to break this cease-fire between minds and machines⁵. LLMs can carry out context-sensitive translations, they can explain humor and jokes including interpreting what humans take to be funny images (Hessel et al., 2023, albeit with notable limitations), can learn from few or a single example or instruction (Brown et al., 2020), or engage on reasoning chains “creatively” (Wei et al., 2023). Moreover, LLM technology and transformer architectures are being applied to multiple sensorimotor modalities both in real-world physical robots through VLA (Vision Language Action) models (Brohan et al., 2023; Collaboration et al., 2023), Q-learning enhanced LLMs (Chebotar et al., 2023), or directly

procedure and conclusions, attributing much of the reported failure to evaluation artefacts and token-budget limits (Lawsen, 2025).

⁵Although some pre-transformer successes were already anticipatory of the progress that AI development was about to suffer. First it was GO, an open-ended, combinatorially explosive game that cannot be played but by intuition in a manner that Go fans consider that is a pure expression of the player's soul. And second, perhaps more importantly, in playing different computer games, using human controls (e.g. first-person visuomotor feedback), and without knowing or encoding the game rules in advance (Schrittwieser et al., 2020).

applied to visual and sensorimotor tasks (Bousmalis et al., 2023; Reed et al., 2022). These expansions of LLMs might succeed out of the text-image bound domain into physically enacting sensorimotor correlations and learning skillful coping with the physical world (Xiang et al., 2023). An approach that brings new generation AI systems much closer to traditionally “AI-skeptic” embodied and situated approaches to mind and cognition (Chemero, 2009; Dreyfus, 1992; Johnson & Lakoff, 2002; Varela et al., 1991). It is, however, too early to judge the success of embodied robotic implementations of new generative technologies. On the contrary, there is a suite of text-interfaced LLMs (ChatGPT, LLaMa, Gemini, Claude, Mistral, DeepSeek, Qwen, etc.), providing first-hand experience for millions of people, defining the way we are relating to LLM technologies and transforming the digital (and non-digital) human environment.

The best way to avoid alienation is not to feed inflationary positions or to join the deflationary ranks, but to find the right ontological categorization for these systems; or, to say it with Simondon (2017), to identify the *mode of existence* of technical systems. Like the case of consciousness or sentience, we cannot leave the answer to “social relationism”; i.e. to a mere social contingent convention on what type of systems deserves which treatment (for a detailed argument against social relationism and AI see Torrance, 2014). A proper understanding of what LLMs *are*, requires delving deep into their concrete structure, operations and coupling with their *milieu* (humans and other machines). Can, and should, we take them for agents? Are they intelligent? If not, what are they? What is the best way to conceptualize them? The answer has important implications in the field of ethics and legal studies (Bertolini & Episcopo, 2022; Clowes et al., 2024; Coeckelbergh, 2021; Fournieret & Yvert, 2020; Mabaso, 2021), but also on the social adoption of these technologies, our collective awareness of their limits and potentialities. We need conceptual resources to organize our experience and interactions with ChatGPTs and their place in our sociotechnical world.

Understanding in some detail how GPT and other LLMs are trained and how they function is of fundamental importance to characterize their “nature”, genuine capabilities and possible implications. In the next section, we provide an explanation of how GPT works, with the goal of contrasting its actual functioning with some ontological attributions, particularly its agentic capacities. A detailed understanding of GPTs functioning will also help characterize its mode of existence and the ways in which it can potentially transform human agency⁶.

⁶ The next section provides a detailed technical overview of how LLMs function, intended to ground the subsequent philosophical analysis in Sects. 4 and 5. Readers already deeply familiar with the architecture and training procedures of transformer models may choose to skim this section or proceed directly to the philosophical arguments beginning in Sect. 4. Note, however, that specific details will be drawn upon in the later sections.

3 How do large language models work?

LLMs are so-called “artificial intelligence” systems (Norvig & Russell, 2021), part of current NLP (Natural Language Processing) technologies, which belong to the family of “machine learning” and the sub-category of “deep learning” systems. They are designed to process and generate “natural” language through a large number (on the order of billions) of processing steps. Transformers (Vaswani et al., 2017), in turn, are one kind of recently very successful type of LLMs, and GPT (Generative Pre-trained Transformer) is a specific type of implementation of Transformer technology (Brown et al., 2020; Radford et al., 2019). Most popular LLMs are delivered as a service by a provider. So, for instance, *ChatGPT*, is a specifically tuned and interfaced version of GPT (and, increasingly, a platform to connect GPT to other tools and to deliver personalized services with GPT technology)⁷.

ChatGPT uses different versions of GPT models to produce human-like text (GPT-4.5, GPT-o3, etc.). It is a computational language processing system designed to generate sequences of words, codes or other data (more recently, images) from an input sequence called “prompt”. Thus, given a prompt, GPT produces the text that would have been more statistically expected on the training data. For example, if the sequence “Elephants don’t play” is entered as a prompt, the ChatGPT offers “Elephants don’t play **chess**” as a response. The system has a heat parameter that increases less likely variations on the output. So, for instance, the system might respond to the original prompt with “Elephants don’t play **video games**” or could simply output “Elephants don’t play.”. This basic functioning is what made so popular the characterization of ChatGPT as simply a complicated auto-complete tool (Floridi, 2023).

However, the simplicity of the general task of optimizing to predict the following word, and its recursive iteration, is the key for the emergence of complex capacities in LLMs. Moreover, optimization alone provides no ground to understand the working of a system, its capabilities and limits, its mode of existence. Appealing to partial aspects of how they operate, Transformers are often qualified as stochastic, probabilistic and statistical (Bender et al., 2021; Chomsky et al., 2023; Floridi, 2023). *Stochasticity* refers to the randomness of how the final output is “selected”. *Probabilistic* is used to indicate that this final decision is taken randomly but on the basis of an assigned probability that is, in turn, allegedly extracted from the *statistical* properties of the training data. Understood on its most generous terms, such descriptions are relatively correct but partial and incomplete. It is possible to imagine a strictly statistical AI that simply computes or extracts conditional probabilities of all possible output tokens given an input stream. But this simply does not work. In fact, there could not be sufficient training data on the universe to make such a machine effective on the basis of pure probabilities or statistics (Wolfram, 2023). And it is not what GPT does. As their name indicates, LLMs create *models* of language. That is, they do not simply store statistical relations or conditional probabilities but instead

⁷ On what follows we shall use the terms “Transformer”, “LLM” and “GPT” and “ChatGPT” almost interchangeably, unless specific reference is provided to the concrete model (e.g. GPT-2, DeepSeek-r1) or to aspects of their interface.

constitute compressed and structured engines that process and transform text input in non-linear, highly interrelated and complex forms.

Before we start with the description of GPTs architecture and processing, it is important to stress that, in general, the processing blocks and procedures are, technically speaking, complex (Bechtel & Richardson, 2010). They do not make “sense” from the point of view of the functional decomposition of the treatment of the input. Certainly not one that a human might understand or guess as a reasonable strategy. Admittedly, there is no theory for how and why this architecture works (Wolfram, 2023). And yet it works, and we have access to the processing architecture of GPT–2 (Radford et al., 2019), and some details of GPT–3 (Brown et al., 2020), to better delimit, without a possibly full understanding, how the system operates.

In what follows, we will provide a detailed explanation of GPT-3 to the best available knowledge, as the prototypical LLM and as an example of generative AI more generally. We have chosen GPT-3 as our foundational case study for several key reasons. First, its release marked a crucial inflection point⁸. It stood for a qualitative leap in generative capability and widespread public adoption. Second, it did set up the industry standard for other models to follow, becoming the paradigmatic example of generative AI. Subsequent models have increased in size, explored more optimization with novel training techniques or, as we shall see, multiple parallel specialized transformers and multimodality. But they largely build upon the core principles of the Generative Pre-trained Transformer architecture that GPT-3 popularized. A detailed understanding of this particular model provides the essential conceptual scaffolding for grasping the entire family of these systems, including their more recent and complex versions, and those developed by other organizations.

3.1 Architecture and processing

Figure 1 illustrates step-by-step the processing of the input text as it goes along the GPT architecture. We explain each step in detail below.

1. Tokenization and encoding: The first operation that takes place as we enter text into a LLM like ChatGPT is *tokenization*. The input stream is chopped into *tokens* (small syllable-like or small word text chunks, including punctuation marks). On average, each token is about 3/4 of a word in English, with a mean of 4 characters per token⁹. Nevertheless, we are going to use the terms “word” and “token” interchangeably. Then each token is encoded numerically. So for example, “Elephants don’t play” is tokenized into five tokens: [Ele, phants, don, ‘t, play]. And then each of them is assigned a predefined number out of the complete vocabulary of 50,257 tokens, and the sequence is converted into an array of numbers: [46439, 53667, 1541, 956, 1514].

⁸ This is so to the extent that the term “GPT-3 moment” has now been popularized, referring to the breakthrough at which a technological advance converts prior hypotheses or intuitions into a clearly recognisable technical success.

⁹ It is possible to work with Tokenizer to understand better this procedure: <https://platform.openai.com/tokenizer>.

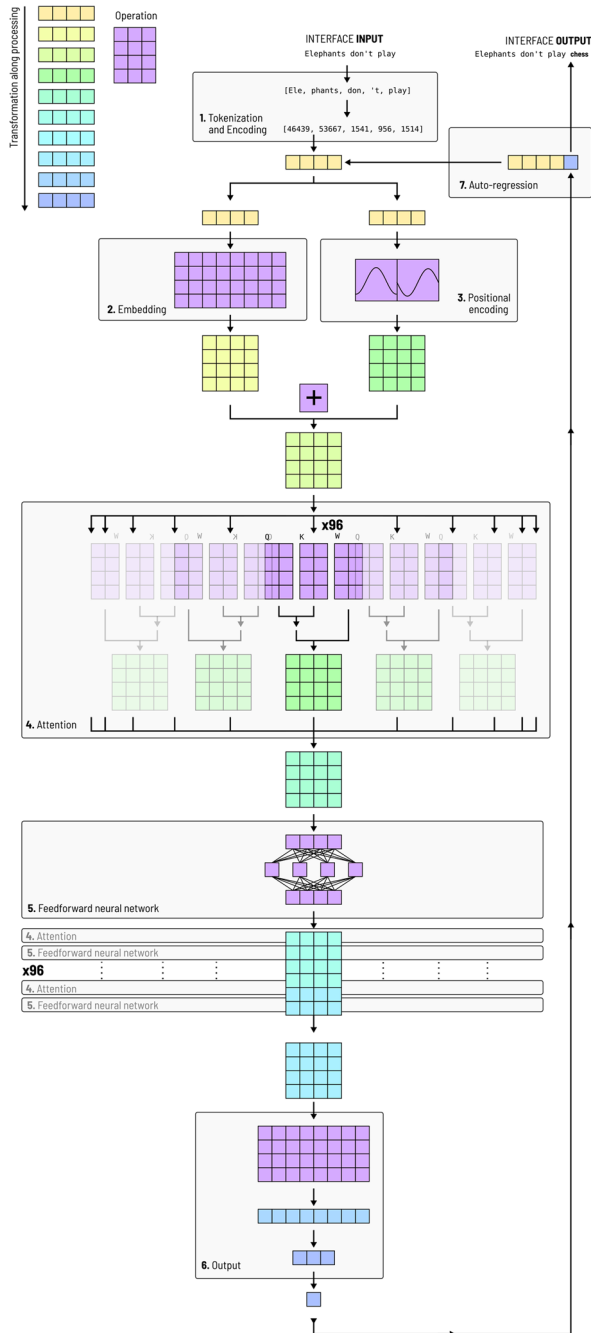


Fig. 1 Schematic of the GPT-3 processing architecture as a standardized reference for LLM functioning. Computation flows from the top of the diagram to the bottom: the input string (shown in yellow) is progressively transformed into intermediate matrices whose colour shifts from yellow to green to blue, culminating in the selection of a single blue output token. Parameter matrices (marked in purple) act on these intermediate matrices (they remain fixed during processing but are configured and updated during

pre-training and training, see sect. 3.2). (1) The input string is tokenised and numerically encoded. (2) Tokens are embedded as 12288-dimensional vectors or "semantic space" defined for a 50257 token (word) vocabulary. (3) A positional encoding is added to each vector to preserve word order. The resulting matrix passes iteratively through 96 sequential transformer blocks, each involving the processing through (4) 96 parallel self-attention heads and (5) a high dimensional feed-forward neural network. (6) A projection back into vocabulary space followed by softmax normalisation yields a probability distribution over the 50257 possible tokens, from which the next token is sampled. (7) The sampled token is added to the original input (returning to step 1), and the procedure repeats auto-regressively until an end-of-sequence token is produced or the maximum length is reached. The text body provides a detailed explanation of each processing step.

2. Embedding: Once tokens are represented numerically, these numbers are mapped into a high dimensional relational space. This process is called embedding. Embedding already implies a huge transformation of the input with previous "knowledge" of how tokens (or words) relate to each other. Some of these relationships can be considered "semantic" or "syntactic" by capturing higher-order relational properties between words. Some dimensions or combinations of dimensions might be thought of as abstract conceptual properties (e.g. being a grammatical subject or being an animal). What defines the conceptual content of each dimension is not an arbitrary label into it, but purely relational "spatial" properties. For instance, animal names will appear close to each other, also grammatical subjects, etc. The embedding space of GPT-3 is of 12,288 dimensions (Brown et al., 2020) and a position within that space is pre-encoded for 50,257 tokens¹⁰ (Radford et al., 2019). One way to understand this (with the risk of anthropomorphizing) is to say that GPT3 has the capacity to situate 50,257 words¹¹ in a 12,288 dimensional "conceptual" space (or along 12,288 "properties"). So, for instance, the tokens "cat" and "tiger" will be close to each other in many dimensions but will be relatively spaced in "size" and "habitat". This means nothing other than the token "cat" being closer to "laptop", "dog" and "watermelon" on a given dimension (that we could interpret as "size") and closer to "Roomba", "television", "sofa" and "living-room" in another (that we might interpret as "habitat").

Embeddings already embody an important part of the "knowledge" of an LLM, and its production is part of the overall training procedure of GPT. The result of applying the embedding function to the input stream is a matrix of $2048 \times 12,288$ that is itself called embedding¹². The number 2048 (for GPT3, the size in current models is much higher) indicates the maximum size of the input stream in tokens. It sets the upper limit of how much "context" (previous conversation, additional information or maximum input provided) can the system handle. Although the input might just

¹⁰ We provide token and matrix size details here with two specific goals: (a) to illustrate the scale of these systems—a point that will turn relevant in Sect. 3.3—and, (b) to track the matrices that are combined at different steps of the process; the size of the matrix is thus provided as a label or identifier. The specific size is, of course, not relevant, but helps track the matrix transformations along the text and in relation to Fig. 1.

¹¹ We will use the terms token and words interchangeably for a more intuitive grasp of the functioning of the system. It is difficult to be strictly rigorous here because the concept of word itself is ill-defined, with a regular human knowing approximately 10k word families (Brysbaert et al., 2016).

¹² It is a regular practice to use the term embedding to name the matrix that will be processed along the whole transformer. But this might lead to confusion. Although the matrix maintains the same size, and the end result will be transposed back into tokens (see latter), the successive operations carried out over this matrix distort its original interpretation so much that we find it confusing to keep calling it "embedding". We should use the term "matrix" instead and identify it by its size.

be of 5 tokens and the resulting matrix $5 \times 12,288$, from now on, for convenience we will assume that the matrix being processed is this $2048 \times 12,288$ maximal sized one.

3. Positional encoding: Not only the set of words composing the input matters, but their order also does. It is not the same to say that “John Searle invented the Chinese Room” than saying that “The Chinese Room invented John Searle”. So each word/token embedding will also be transformed to incorporate positional information. The positional encoding in the form of a unique sine and cosine function output is added to the word embedding. A wave signature is added to each embedding array that is unique to a specific position and can be exploited by later processing to identify the position of that token on the original input stream. This produces a new matrix of $2048 \times 12,288$ with all the 2048 input tokens in one dimension and their embedding+position on the other 12,288. The combination of embedding and positional encoding will now be processed through a sequence of processing blocks, like a factory line. Each block consists of a set of operations that include primarily: attention, addition and normalization, and feed-forward neural network processing. GPT-3 transforms the input matrix through 96 such blocks.

4. Attention mechanism: This is the most innovative of all the steps on the LLM revolution and characterizes transformers as a specific type of LLMs (Vaswani et al., 2017). Attention layers have allowed an increase in LLM size and efficiency due to their capacity to parallelize processing during learning and execution time. They permit to explore a wide range of correlation dependencies over the input data, in a highly scalable manner; improving upon other architectures aimed at processing relationships between text elements in an input (e.g. retaining a working memory of the past words in a paragraph), like Recurrent Neural Networks or Long-Short Term Memory Networks.

Attention mechanisms basically compute how important are some tokens in a stream (and how other tokens can be ignored) but also how important are the relationships between tokens in the input stream. This might pick out short and long distance relationships, chopping the input stream into different chunks. Some of these attentional relationships might capture grammatical connections, such as a verb whose subject is far back in a sentence. Others might capture instructional (e.g. the relationships between different steps of a recipe) or narrative structures (like the unfolding of plot and the connection between characters through time). This is transiently expressed as a matrix in which all the tokens are valued in relation to all other tokens relating “everywhere, all at once”, in what would horrify Bergson as a geometrization of duration. By computing in parallel 96 attention mappings of this kind, transformers avoid the computational bottleneck of recurrent sequential processing. In short, attention mechanisms make it possible to be sensitive to different contextual scales. GPT-3 processes 96 attention heads¹³, that means that it processes (and later combines) 96 different ways of relating tokens of the input sequence to each other. What exactly each of these heads “really pay attention to” is partially unknown.

¹³Not to be confused with the 96 blocks. Attention heads run in parallel inside each block. Block processing takes place in a sequence, the matrix that results from the transformations of block-*n* are the input to block-*n*+1. Attention heads process 96 copies of the input matrix in parallel, and then all 96 are added and normalized into a single matrix that is then further processed. See Fig. 1.

5. FeedForward network: The next step involves passing the matrix through a FeedForward Neural Network (FFNN) and it involves an important expense on the dimensionality of the processing and the non-linear interaction between all the components of the matrix; relating “everything, all at once”¹⁴. The feedforward network consists of 3 layers. The input layer is the matrix itself, the hidden layer expands its dimensionality and the output layer reduces it back to the original size. All the nodes of the first layer are connected to all the second, and all the second to the third (but not between themselves nor backward, thus the name FeedForward). The connections are weighted, so that different relationships between value projections can have different weights and amplify or reduce the value of each signal into the next layer. This is then processed by the nodes of the next layer through a nonlinear function. In principle, FFNNs can compute any function (Siegelmann & Sontag, 1995). In this case, they could be thought of as a computer inside a computer that could simulate different types of programs¹⁵, with the benefit of being programmable in an unsupervised manner (see training section below). The weights and the parameters of the nonlinear function have traditionally been understood as the “place” where “knowledge” is encoded (Churchland, 1990; Rumelhart et al., 1987). So, for instance, the fact that GPT responds with “chess” to the sentence “Elephants don’t play”, instead of simply “.” or “basketball” is not something to be found on the original embedding, where certainly “chess” is an option close to “play” but certainly not the closest. Traces of Brook’s famous paper “Elephants don’t play chess”, its poetic “value”, and other contextual elements (e.g. talking about AI and the notable role that chess played in its history) can explain the final output.

Attention takes over 30% of parameters of the model and FF about 70% in the largest GPT-3 175 Billion parameter model (Huben, 2023). After embeddings, positions and attentional processing has taken place, FFNN processing “elaborates” relations between tokens applying to them the knowledge that was acquired through the training process. But arrays of the resulting matrix can hardly be understood as directly relating to the original tokens anymore. Less so in subsequent transformations, since the matrix will now be the input to another block that starts again with its 96 attention heads (different to those of the previous block) and its FFNN processing that has also specific parameters (weights and biases) in each block. At the end of the process, the original matrix is severely transformed on its values and is ready to be finally transformed into the output.

6. Output: the $2048 \times 12,228$ matrix that resulted from the processing of previous blocks needs now to be converted into a single next token for the original input array. Recall that the original embedding projected a vocabulary of 50,257 words (tokens) into a 12,228 dimensional space. A $12,288 \times 50,257$ projection matrix (which is a transpose of the embedding) now transforms the processed matrix into a score for each token of the vocabulary. The top-k highest scored tokens are separated, and a

¹⁴This is the most unknown or unexplainable part of the transformations that the input suffers. The operations are simple and vaguely inspired on how natural neuronal networks function. But what exactly are the structural changes that take place and what they correspond to in terms of humanly explainable linguistic or cognitive operations are fundamentally unknown. And might inevitably remain so.

¹⁵ Although they are not technically “Turing complete” due to deficits in recursion and memory.

softmax algorithm simply transforms each token punctuation into a normalized probability that is proportional to its score. Then a final token is selected according to these probabilities. Visually, this can be likened to a roulette wheel, where each segment's size is proportional to the token's assigned probability. The selection process mimics the spinning of this wheel, with the chosen segment indicating the next token in the sequence.

7. Auto-regression: The above sequence of operations is repeated again and again, adding a new token to the end of the string (e.g. a new word to the sentence) until the maximum number of allowed tokens is reached or, most commonly, an end-of-sequence token is produced by GPT (a kind of “halt” token that is interpreted as a stop). It is possible to continue the process by reintroducing the input again and adding some more text (like when we add a response to the conversation). Although the fact that GPT'S “intelligence” is often displayed when it stops, the most relevant aspect of auto-regression is the type of “externalized” feedback that it provides for the system. And this is an essential part of its functioning. Note that in no step of the architecture so far did ChatGPT store any information, there is no “internal” state, no memory. It is, in a sense, a purely reactive system¹⁶. It is through auto-regression that it does compensate for it, in a manner that will become very important when addressing the agentic capacities of GPT.

The term transformer was originally chosen to depict the sequential transformation of the input matrix into an output without recurrence, with the task of translations as a key component (Vaswani et al., 2017). It was later applied to other tasks (like summarization) and finally discovered that a large enough transformer could perform very well generally across tasks. And also that it could “learn” what to do simply by direct instruction with none or few examples of what was asked to do (Brown et al., 2020; Radford et al., 2019). This is when the concept of prompt takes significance. The power of LLM transformers is so general and unspecific that it is open to be prompted to unfold in different directions: summarization, translation, correction, explanation, conversation, expansion of key ideas, development of outline strategies, etc. The “magic”, so to speak, that sustains these capacities, lies on the parameters of the system, the embedding, attention and FFNN matrices (colored purple on Fig. 1) that operate on the input matrix.

We know very little of GPT-4 and subsequent OpenAI models. Despite its name and the original goals of the company soon after GPT-3 it moved into secrecy, releasing very little information of new models, other than some raw details and benchmarking results (OpenAI, 2023). It is speculated that one of the greatest architectural innovations of GPT-4 over its predecessors was the inclusion of Mixture of Experts. This is a technique of splitting the network into specialized subnetworks and training it to learn to distribute processing to them (Shazeer et al., 2017). We directly know about the effectiveness of this technique, because it has been effectively implemented on open source or and more transparent LLMs that match GPT4 benchmarking performance like Mixtral (Jiang et al., 2023), LLaMa (Meta AI, 2025), KIMI K2 (Moonshot, 2025/2025) or DeepSeek (DeepSeek-AI et al., 2025). The other

¹⁶You can explore this yourself by asking GPT to imagine a number or retain something secretly and perform operations on it and the like.

important innovation is multimodality. Tokenization has nowadays turned multimodal. Images are broken into a grid of patches (e.g., 16×16 pixels) and each patch treated like a “word” (a token), thus creating a sequence of “image tokens” that can be fed into the main Transformer (Dosovitskiy et al., 2021). The same goes for audio waveforms. Once encoded, all these different tokens (text, image, audio) co-exist in the shared, high-dimensional vector space, where relationships and patterns can be learned between modalities¹⁷. Although functionally relevant, these innovations do not imply a radical change from the explanation we just provided. The architecture remains fundamentally the same.

3.2 Training

GPT and other LLM configuration is typically carried out in various stages. The first is, somewhat paradoxically, called “pre-training” but constitutes the main training (understood as the process by which one improves or acquires new capacities). During this process, the parameters of each processing transformation just described gradually change until a given level of accuracy is reached, pre-training ends, and they remain fixed until new training procedures start. Then comes fine-tuning, with two basic stages: task specific fitting and reinforcement learning. Finally, prompt learning is often used, which is more of an instructional form of directing the system.

3.2.1 Pre-training

Explaining first the way of functioning of the whole architecture, as we just did, is essential to understand training. Contrary to other approaches, each processing block is not trained in isolation to perform a specific task (e.g. 1st grammatically articulate the input, then build a general abstract representation, next, carry out inferences and take an output decision), but the entire system is trained at once, through *back-propagation* (Rumelhart et al., 1986).

The basic mechanism is simple: the system is initialized with random parameters. Next, a chunk of input (e.g. the beginning of a sentence) is chosen among a training dataset. It is then processed as we explained above. This is called a forward pass. This pass finishes when the system provides the result array: that which indicates the probabilities of all the words to be the next one (the step before selecting the final output). The result will be nonsense at the beginning. For example, to the input “Elephants don’t play” the highest probability of the result array could be “purple”, followed by “Fodor”, “misuse”, “chain”, “cat”, etc. Now, this is compared to the correct result: an array that gives 0 probability to all the words except 1 for “chess”. But “chess” might be very down on the assigned probabilities. Yet, it is now possible to compute

¹⁷ Multimodal generation is more complicated. Instead of recursively adding tokens to a string, like language models do, image, sound, or video synthesis typically relies on generative *diffusion models* (Ho et al., 2020). They work on what can intuitively be understood as a kind of “internalised sketch pad”. The model starts with a tensor of pure random noise and iteratively refines it by subtracting the noise until a clean image emerges. In multimodal text-to-image systems, this “denoising” process is conditioned on the text prompt, steering the generation towards an output that is “semantically” aligned with the user’s description.

an error (or loss): the difference between the assigned probabilities on the result array and the target one.

Next, this loss will be backpropagated through the network (the backward pass). By means of an optimization algorithm, small changes are made all throughout the whole network in the direction of minimizing this error: The algorithm calculates the response to the question “what change should I do to this parameter so that the resulting output reduces the error?” and makes the change accordingly, for each parameter on each block, backwards.

This process is iterated once and again, until the forward process produces no or little errors. All three major components of the LLMs are trained in this way: embeddings, attention mechanisms and feed-forward networks. Although the overall procedure is locally relatively simple, the amount of little changes is vast and the effect is the performance capacity we can witness today. The computational cost of training GPT-3 was 3.14×10^{23} FLOPs, that is, 314 sextillion floating-point operations (Brown et al., 2020 Appendix D).

The system so far is considered a raw *foundation model*: it can process text generally and can be put to work on a number of tasks already or be further trained to improve performance on specific types of tasks. The training process, so far, is considered unsupervised, nothing other than the next-word-prediction is used to train the model.

3.2.2 Fine-tuning and reinforcement learning

Additional training procedures are used to fine-tune the transformer for specific tasks, like summarization, translation, or conversation. This time the instruction (e.g. summarize) and the task input (e.g. a whole Wikipedia article) are provided, and the system is trained with back-propagation to match a model output (e.g. Wikipedia’s summary entry for that article), instead of just the next token. This is considered *supervised learning*, no human intervenes yet, but the task is not simply to “guess” the following word but to match a specific target goal, pairs of input and target-output are required to complete this training.

Transformers are usually further trained to include *Reinforcement Learning with Human Feedback* or RFHF (Ziegler et al., 2020). The pre-trained and fine-tuned LLM is let to interact with humans. Then, based on how humans have positively or negatively evaluated the output of the model, it is trained to produce outputs that are more likely to be positively rated or less likely to be negatively valued; according to the past corrections made by human interactors. This is where the system is often trained on ethical or moral values, together with a number of other quality checks.

Finally, we have *one-shot or few-shot learning* procedures that operate basically at the prompt level, providing examples or specific instructions that the LLMs take as input to produce new examples or follow the instructions provided (Brown et al., 2020).

3.2.3 Self-improvement and AI-supervised training

In order to overcome the limitations of LLM to solve complex problems (e.g. in maths or coding), new innovative training techniques have been introduced. Interestingly, they often involve the use of AI generated feedback, data, or supervision. In Process-supervised Reward Models (PRM), for example a specific artificial model, the PRM, is first trained using a huge database of reasoning paths where every step has been labeled by humans as correct or incorrect. The first goal here is to train the PRM to evaluate good reasoning, assigning rewards to each individual reasoning step. The LLM is then trained by the PRM, acting as a “teacher” via reinforcement learning (RL) to produce novel reasoning chains given new prompts (Lightman et al., 2023). This would be equivalent to teaching first a machine to identify good steps on cooking recipes, and then using it to teach a cook to generate good recipes by judging each step to achieve the final (prompted) request.

Other methods involve, instead of reinforcing specific steps on a reasoning chain, to differentially reward different chains of thought according to their relative quality (Shao et al., 2024). In this case, the student is left to generate different recipes and is rewarded according to how good each recipe would comparatively result to a culinary critic (who does not really know *how* to cook but has been trained to develop a good taste). This is the formula behind the success of DeepSeek-R1 (DeepSeek-AI et al., 2025). A further innovation from this company involved distillation, where a large, powerful “teacher” model is used to generate the vast synthetic dataset of high-quality reasoning paths and ranked-choice answers¹⁸. The smaller, more efficient “student” model is then tuned on this machine-generated corpus (similar to the pre-training phase). Following the metaphor, once a good cook is trained, it is possible to ask it to generate millions of new recipes and to train a smaller (cost-efficient) model with them, overcoming the limited amount of recipes in available cookbooks.

These AI-supervised procedures can be applied not only to improve reasoning, but also to “induce” ethical guidelines. This is the case of *Constitutional AI*, where the model learns to align its behavior with a set of written principles, a “constitution”, defined by humans, but does so without direct human feedback. The process involves an initial phase where the model is prompted to critique and revise its own responses according to the constitutional principles. The system is then tuned to predict the next token of those self-revised responses. In a second phase, a teacher preference model is trained on the constitutional principles and then RL with AI feedback is applied to the LLM (Bai et al., 2022).

3.3 Anthropomorphising GPT

Calls to avoid anthropomorphizing GPT are recurrent (Bender et al., 2021; Butkus, 2020; Coeckelbergh, 2021; Jebari & Lundborg, 2021; Kubes & Reinhardt, 2022; Shardlow & Przybyła, 2023). But anthropomorphizing here only referred to pro-

¹⁸ The DeepSeek team generated one order of magnitude more reasoning chains artificially for distillation (about 800 K) than the original dataset presumably used by the OpenAI team with 75 K examples (Lightman et al., 2023).

jecting human qualities, particularly cognitive or emotional ones, to the machine. Something that is perceived as a risky strategy, since understanding (or experiencing) the interaction with ChatGPT through the human or intentional stance (Dennett, 1989) as if it truly had genuine human capacities, would make us falsely attribute a set of properties it certainly lacks. Properties that are essential to the human social world-making: commitment, trust, responsibility, empathy, etc. Important as it is, the emphasis of this type of anthropomorphization shadows other important forms. There are at least two more types of anthropomorphization that are relevant to understand GPT. And their analysis is perhaps more revealing of its mode of existence than the attribution of mental properties to the system. The first such type of anthropomorphizing is the way in which the training corpus and procedures shape the machine as a human. The second is the inverse process of trying to bring to the human scale and capacities the internal workings of the system. We shall attend to both in this section in a somewhat combined manner.

Regarding the processing, according to Kaplan et al. (2020), we can roughly approximate the computational cost of processing a single token to be directly proportional to the number of parameters of the model. GPT-3 having 175 Billion parameters, the computing cost of writing a 250 token summary of a 1750 token essay could have the approximate cost of $2000 \times 175 \text{ Billion} = 350 \text{ trillion FLOPs}$ (floating point operations)¹⁹. Carried out by a human, as an experientially graspable task, each FLOP could be approximated as equivalent to a multiplication between two 5-digit numbers. Assuming such an operation could take about 10 s to be completed by an expert or well-trained human being²⁰, it would take around 500 million years of human labor, working 40 h a week, to process that prompt²¹. That means that John Searle would have to live and die *a few million times* before he could output even the first symbol from his Chinese room. Intuitions that once worked at a certain scale (like the Chinese room experiment) might not necessarily be trusted at many orders of magnitude higher scales.

Regarding the training aspect, we know little of GPT-4 but GPT-3 was trained with 570 GB of text data, about 300 billion tokens according to their own creators (Brown et al., 2020), that is approximately 200 Billion words. Thus, the training data for ChatGPT is equivalent to about 2 million books. A volume so vast that it would take a human being more than 500 years to read through it all, assuming they dedicated 8 h a day, every day, reading at a speed of 200 words per minute. To put this into further perspective, if an average person reads roughly 500 books in their lifetime, the amount of data ChatGPT has been trained on is comparable to the combined reading of 4000 lifetimes. But if we were to humanly compute all the backpropagation process of 314 sextillion FLOPs, that would take an expert human 4.19×10^{17} years

¹⁹Although the recursive nature of the output processing could indicate an exponential growth of this cost, it is effectively reduced by not re-processing the whole input again (see this discussion for a more detailed explanation Tunstall, 2022).

²⁰It actually took one of us a few minutes to complete it!

²¹At the same time, it is worth noting that the human brain, at a subconscious, subpersonal level of activity, can carry out the equivalent of this 350 trillion FLOPs in about one second or less (Carlsmith, 2020).

to compute, which is almost 7 orders of magnitude (30,386,783 times) the age of the universe.

Let's now focus on the inverse analysis of how ChatGPT is already anthropomorphized by the training data, including the biases, themes, styles and poetic tendencies that are present in them. And not less importantly, by all the fine-tuning and human reinforcement learning. To say it differently, ChatGPT has no way of organizing tokens around mothers, mice, or forests other than that provided by human traces on texts. It is, thus, not surprising that we can anthropomorphize it. It already is. And it is so in a manner that cannot be fully grasped. Unlike a mannequin that we can touch and verify that its shape is indeed human, but whose functioning is nothing more than that of a piece of inert plastic. We can not even bring the complexity of the concrete functioning of GPT down to a graspable human scale. Its intensive and enormous training procedure, its gigantic internal structure and its vast mode of operation lies beyond the human scale of understanding²². However, its internal and behavioral functioning can be generally (if not specifically) sufficiently understood so as to determine constraints to its mode of existence. And we can properly ground why and how we can avoid ontological anthropomorphization. Agency being a pivotal, often anthropomorphized, category for settling the ontological status of LLMs.

3.4 Towards LLM based agents

At a first sight, nothing in this architecture qualifies properly as agency. Not even for the most optimistic or naive engineers. The system is fully driven by the prompt and directly steered when output completion has taken place by a new prompt. Moreover, the system has no internal states, no (internal) memory, no potential desires, goals, or purposes. When operating (after training is completed), not even a trace of what is processed is left within the system; except for the history of outputs that is continuously fed-back into the system auto-regressively. In a sense, GPT operates like Leonard Shelby, the protagonist of Christopher Nolan's celebrated film *Memento* (2001). Devoid of the capacity to create new memories (yet able to use its knowledge), Leonard externalizes instructions (goals, instrumental steps, etc.) and contextual information (pictures, notes, etc.) to regain the agency that he lost due to his amnesia.

The lack of agentic capacities of the raw GPT is apparent in the type of digital embodiments that AI engineers are providing to enhance GPT and develop so-called "autonomous GPT agents" (Andreas, 2022; Huang et al., 2024; Wang et al., 2023; Weng, 2023; Xi et al., 2023; Zhou et al., 2023a, b). March to June 2023 saw a rapid increase of projects trying to deploy digital agents based on GPT and other LLMs: AutoGPT (Significant Gravitas, 2023/2023), AutoGen (Wu et al., 2023/2023), DemoGPT (Ünsal, 2023/2023), SuperAGI (admin_sagi, 2023), MiniAGI (Mueller, 2023/2023). A number of initiatives have followed that promise to deliver fully operational agents for programming (Wu, 2024; Yang et al., 2024) and tech giants seem to be betting on LLM-driven agents to make generative-AI services profitable (Holmes, 2024; Knight, 2014).

²²So does a bacterial cell or the global economy, by the way; not to mention the human mind and the extended socio-cultural scaffolding.

There are 5 kinds of LLM enhancement strategies that are being developed to move from ChatBots to the so-called “agents”: (a) extended memory systems, (b) planning strategies, (c) reflexive evaluations, (d) the use of tools, and (e) multi-agent interactions. These strategies are most often implemented in combination, but it is, nevertheless, possible to differentiate them.

Extended memory frameworks (like Langchain) make it possible for transformers to temporarily extend and sediment autoregressive dynamics (e.g. rewriting and organizing summaries of past input context to increase its memory and better focus it on specific task-goals). Sometimes such extensions are not different from our practice of externalizing memory in a notebook, writing To-Do lists and offloading planning structuring in a bullet point document.

Planning strategies involve prompting the transformer to split a specific goal or task into sub-operations that can then perform in sequence. This can be achieved through various techniques, the most known of which is the so-called Chain of Thought or CoT (Wei et al., 2023). CoT is implemented by crafting prompts that encourage the model to “think aloud”. Instead of trying to answer directly to a given question, or to accomplish a task, the LLM is first prompted to explicitly write down how it will plan to do it and then follow its own plan. This technique has been shown to enhance LLM reasoning and planning abilities. More sophisticated methods, like Tree of Thought (Yao et al., 2023; A. Zhou et al., 2023a, b) involve combining a tree-like decomposition of a variety of plans with LLMs capacity to reflexively evaluate the adequacy of each potential plan (which brings us to the next point).

Reflexive evaluation procedures as simple as asking the transformer to reflect on the previous output and correct existing mistakes have been shown to dramatically increase performance (Madaan et al., 2023b), also to provide some degree of self-guidance on the completion of the decomposed sub-tasks. The generative and creative capacity of LLMs to deliver execution plans is often combined by using LLMs to automatically evaluate them and to distill a more consistent strategy. Both planning and reflexive evaluation are often now built-into “reasoning” LLMs (as explained in Sect. 3.2.3).

Use of tools LLM can be connected to a wide variety of tools (Mialon et al., 2023; OpenAI, 2025b; Schick et al., 2023; Shen & Yang, 2025), from programming consoles like Python, to web-browsers, search engines, and, more generally APIs (Application Programming Interfaces) that make possible to interact with digital services through instructions (rather than visuomotor interfaces). These “tools” define the “bodies”, interfaces and environments of LLM powered “agents”. In November 2024, Anthropic launched the Model Context Protocol to define such tools and interfaces for LLMs, and it has rapidly expanded as a standard to let LLMs access different services (Anthropic, 2024; Hou et al., 2025).

Multi-agent interactions Finally, in order to overcome memory limitations, lack of consistency or repeated failure, multi-agent approaches are used, which involves interacting, evaluating and selecting results from other agents (Li et al., 2024). Increasing successful task completion through collective agency is frequently achieved by combining many of the techniques explained above, like self-organizing Tree of Agents strategies (Chen et al., 2024).

Despite the increasing enthusiasm on the potential of LLMs to provide solid foundations for digital agents, strong limitations have already surfaced: LLMs do not seem to be much better discriminating than they are generating plans (D. Jiang et al., 2024a, b), they are very limited on their capacity to develop complex plans (Kambhampati, 2023), perhaps because it is still very hard (Wu et al., 2024) for LLMs to integrate future tokens on their current processing; which is a fundamental way in which humans plan.

In Chat scenarios, human intervention can continuously stir the conversation, discard hallucinations, or ignore wrong answers. Agentic scenarios are different. The human presence in the conversational domain makes the coupled LLM-human system much more fault-tolerant. But when humans are out of the loop, errors can accumulate catastrophically. Think of the cumulative effects of hallucinations or mistakes on making a cake: a mixture of eggs shells, salt, and flour could end up in the fridge instead of the oven, despite a 98% accuracy on the design and execution of the recipe. It is thus no surprise that, unlike benchmarks directed at measuring linguistic capacities, intelligence or knowledge, LLMs still score far behind humans in current agentic benchmarks (Liu et al., 2023; Valmeekam et al., 2023; Xie et al., 2024, 2025). It is therefore still early to judge whether they can at least operate “as if” they were genuine agents. This remains an open empirical issue. Meanwhile, it is possible and necessary to explore how existing transformer architectures meet the requirements for agency identified at a more fundamental level than that of pure performance.

4 LLMs are not (autonomous) agents

We have seen how LLMs based on transformer architecture internally operate and how their capacities have been expanded with a series of additions to the foundational trained models. Moving below the surface of performance-level measurement to characterize agency requires a certain commitment to theoretical or philosophical frameworks. It is essential to specify the nature of actions, purpose, and cognitive properties, and to relate them to the underlying generative mechanisms. The detailed technical account of LLMs’ architecture and agent-like extensions will turn out critical to assess their ontological status. In this section, we first approach the assessment from the point of view of computational representationalism (from which LLMs can comfortably be characterized as agents). We then move to alternative, so called, 4E frameworks, whose requirements severely problematize agency attribution to LLMs.

From the philosophical perspectives that have given credit and have contributed to the AI research program, it is difficult to rule out genuinely agentic capacities from ChatGPT-like systems. Representational computationalism is one such approach (Carruthers, 2006; Newell, 1980; Putnam, 1965). It is a type of functionalism that defines mental properties (intelligence, knowledge, learning, or agency) in terms of the input-output functional (internal transition) states of a system representing states of affairs of the environment. The essential feature of the mind is the capacity to reason or to draw inferences upon representations of the world; i.e., information processing. For instance, you take the umbrella because you just read it will be raining today, and you know that umbrellas are a good way to cover yourself from the rain.

According to representational functionalism, this is the kind of inference that is characteristic of mental processes. And, LLMs are well capable to make such inferences. Moreover, their internal states reasonably approximate the world. Being models of language, and being trained on a huge amount of text, to the extent that all these training data can be squeezed to provide a model of the world, LLMs, are also models of the world (Kadavath et al., 2022; Li et al., 2023a, b; Yildirim & Paul, 2024) including other agents (Andreas, 2022).²³

Some authors have gone even further, proposing that all reality is informational and agency is the ability to act upon and be affected by the (informational) environment. Agency is thus not limited to human beings but can also apply to artificial entities such as robots, software programs, and AI systems:

“These new agents already share the same ontology with their environment and can operate in it with much more freedom and control. We (shall) delegate or outsource to artificial agents memories, decisions, routine tasks and other activities in ways that will be increasingly integrated with us and with our understanding of what it means to be an agent.” (Floridi, 2007, p. 62).

From this perspective, LLMs could be considered agents perfectly embedded in the infosphere. In fact, Floridi has recently contemplated this possibility and the problems and challenges involved, concluding that GPT-like AI systems are “agency without intelligence” (Floridi, 2023; Floridi & Chiriatti, 2020). His category of agency can be understood as depending on two key components: capability and autonomy. Capability refers to an agent’s ability to perform a certain action or set of actions, while autonomy (for Floridi, and much of AI engineering) refers to an agent’s ability to act independently, without being controlled or directed externally. In short, this theoretical framework can be summarized as follows: (1) all entities are informational, (2) some (informational) entities are agents, and (3) agents are entities that perform actions independently of one another.

However, this characterization of agency might be too liberal. Agents are characterized by their capacity to carry out actions, as distinct from mere events or mechanically caused states of affairs²⁴. Actions, unlike (other) events, are not merely occurrences in the world (informational or otherwise); they are processes imbued with intentionality and purpose. This distinction becomes evident when comparing the experience of intentionally reaching for a bike, an action, with being inadvertently pushed towards it by the wind, an event. The former is characterized by a sense

²³ According to this view, having no “real” contact with the world is no fundamental obstacle. Certainly (some) LLMs have no vision capabilities to see if it is raining right now, but nor do you, when you read in the newspaper or your favorite weather-app that it is about to rain. The interface of information reception does not affect the nature of the inference that it is appropriate to bring the umbrella with you. In turn, a LLM, without a robotic body, could not itself complete the action to take the umbrella, but it could perfectly command you to do so (by means of a text message) or could signal the cars’ top window’s controller to close it. The nature of the cognitive process of making the right inference according to the right knowledge of the world is indifferent to the mediation of the input or output. In this sense, LLMs could be considered full-blown cognitive agents with more or less sensory and motor capabilities.

²⁴ Even when it is considered that actions are caused by events (Davidson, 1980), these are of a very special kind: reasons, beliefs, desires, etc.

of directiveness and intention, elements central to the phenomenology of agency we experience and recognize every day. Any output of a system (computing machine or otherwise) does not automatically qualify as an action.

Many have questioned the adequacy of informationalist and computationalist approaches to capture and explain cognitive and agentic capacities. In this critique converge theoretical contributions from different fields: phenomenology (Gallagher, 2017; Merleau-Ponty, 1944), philosophical and theoretical biology (Jonas, 1966; Maturana & Varela, 1980; Moreno & Mossio, 2015), philosophy of mind (Searle, 1980; Noë, 2004; Hutto & Myin, 2012; Thompson, 2010) empirical contributions from the psychology of perception (Gibson, 1979; Heras-Escribano, 2019; Reed, 1996) or conceptual development (Lakoff & Johnson, 1980), large-scale neuroscience (Buzsaki, 2006; Freeman, 2001), and methodological contributions from complex dynamical systems' theory (Barandiaran & Moreno, 2006; Favela, 2020; Port & Gelder, 1995). These and other criticisms have resulted in a family of alternative approaches that are often labeled under the term 4E-cognition (Gallagher, 2023); standing for embodied, extended, enactive and ecological.

Within these, the approach outlined by Barandiaran et al. (2009) allows for the comparison of natural agency with the operations of LLMs (or any other system)²⁵. They start by reviewing different available definitions of agency (from software engineering to robotics, from philosophy to psychology) to bring together a surface description of what these definitions have in common: “a system doing something by itself according to some goals or norms”. They spell out what this commonality entails, identifying 3 necessary and sufficient conditions for agency: individuality, normativity and interactional asymmetry. First, an autonomous agency requires that a system be self-individuated. Second, the self-individuation process defines a set of norms (of viability) and, third, according to these norms, the system asymmetrically regulates its coupling with the environment (thus becoming the source of an action). In sum, from their perspective, an agent system is an autonomous organization capable of adaptively regulating its coupling with the environment in order to sustain itself according to the rules set by its own conditions of viability (Barandiaran et al., 2009, p. 376).

A bacterium moving up a sugar gradient (Berg, 2004) is a widely accepted paradigmatic example of agency that satisfies the definition (Barandiaran & Egbert, 2014). First, the bacterium is in a continuous process of individuation and self-distinction. Metabolism produces the components of the reaction networks constituting the agent, together with a membrane that separates the system from its environment. This self-production in turn determines which aspects of the environment are relevant, normatively valued (good or bad) from the very constitution of the system: some chemical compounds are essential nutrients for its self-maintenance (good), some others are poisonous compounds that degrade the membrane or the metabolic network (bad). Finally, the agent modulates its coupling with the environment by moving up or down a sucrose (positively valued nutrient) gradient and absorbing sugar molecules across

²⁵ For a similar approach that identifies requirements and limitations of AI systems to meet enactive standards for cognition see Froese and Ziemke (2009).

the membrane. The whole combination of self-individuation, norm generation, and adaptive regulation constitutes the agentic nature of the bacterium's behavior.

Processes of individuation and normative regulation need not happen at the metabolic scale exclusively. Mental or sensorimotor life can also ground agentic capacities (Barandiaran, 2007, 2008; Di Paolo et al., 2017), bringing it closer to our own experience of intentional agency. Not only do we experience our living body (our biological agency), but also our actions in the world are guided by goals and intentions that transcend mere biological values. This is so because a new level of autonomous organization emerges through the neural mediation of behavior. The individuation process is constituted by a self-sustaining network of sensorimotor schemes (e.g. habits) in continuous development²⁶.

Our experience of agency stems from the fact that we are a mesh of habits that, through specific actions, asserts its own identity. We shape ourselves as behaving systems by acting. A psychological and cognitive identity develops through activity dependent plasticity, organizing brain, body, and environment. And the norms that emerge from this identity direct my behavior. I identify myself as a philosopher, I want to make a good contribution analyzing GPT, I struggle to write these lines correctly. The goal of a specific task is nested into a network of interests and plans that ultimately rest on preserving our identity (Barandiaran & Rama, 2025).

Yet I can be coupled with my environment in many ways that, despite involving myself and my norms or goals, do not qualify as action. Someone else, chance, or simple physical constraints, might move or prevent certain moves so as to contribute to my own norms, e.g., a nurse in a hospital taking care of me. But the result of this behavior would not qualify as agency yet (in fact, I am a patient in a hospital, not an agent). It is not until the source of my behavior lies asymmetrically laden to my psychological or sensorimotor capacities and my sensitivity to my self-generated norms, that my coupling to my environment can properly be called an *action*.

Does ChatGPT meet the three criteria for agency? Let's analyze them one by one. The first condition, *individuality*, requires that a system self-produces or at least self-sustains, distinguishing itself from an environment it co-defines. LLM's existence and maintenance, however, are reliant on external human intervention and tools, and it operates within a predetermined environment. This diverges from the self-individuation process essential for autonomous agency. Note that it is difficult to tell from the agentic perspective what is the system in GPT. On the one hand, there is the input, transformed into a matrix and processed along the complex network of operations described in Sect. 3. On the other is the transformer that is nothing but a set of operator blocks, blindly transforming the input and the resulting matrices. Underneath is the hardware, whose operations are indifferent to the type of processing taking place, or even to the fact that no processing takes place.

When considering *normativity*, an agent system is inherently at risk of degradation without specific actions, and sensitivity to these viability conditions that emerge from

²⁶For more representation minded approaches, this can also be conceived as a network of beliefs whose main behavioral manifestation is the growth and maintenance of the network. Free energy and active inference approaches are relatively consistent with this view and several parallels have been drawn with enactivism (Clark, 2013; Kirchhoff et al., 2018; Seth, 2021), although severe objections to identify both theories have also been drawn (Aguilera et al., 2022; Di Paolo et al., 2022; Nave, 2025; Raja et al., 2021).

its process of individuation is essential for normative behavior. ChatGPT, however, operates without such precariousness and does not autonomously establish its goals. Its functions are programmed externally, driven by an error or loss function that is completely extrinsic to the operations of the system. This absence of intrinsic normativity is evident in ChatGPT's inability to recognize failure autonomously, often leading to repetitive or non-productive responses. A counterintuitive property of LLM is that they are typically capable, retrospectively, if explicitly asked, to identify "errors" on their previous operations (Madaan et al., 2023b). And yet the absence of implicit normativity is apparent in that, given their autoregressive character, they never "realize" their mistake unless prompted to do so.

Instead of *purposeful* action (behavior that is imbued with purpose all along its unfolding) transformers are somehow partially *purpose-structured* as a result of the pre-training on the massive purpose-expressive texts, and *purpose-bounded*, that is, statistically falling within "humanly interpretable normative" bounds, as a result of fine-tuning and supervised learning procedures. Moreover, training procedures used to shape autoregressive loops as "think aloud" strategies for so-called "reasoning models" endow them with purpose-structured and purpose-bounded generative procedural capacities²⁷. However, unlike living agents, ChatGPT lacks the essential concern or awareness for goal achievement that converts an operation into a purposeful act.

The concept of *interactional asymmetry* underscores that an agent system is the originator of its actions, modulating its interaction with the environment becoming the *source* of the action. ChatGPT, however, shows a fundamentally reactive nature (it has no internal state). It responds passively to external inputs (prompts) and, as we just identified, lacks self-defined norms to guide its interaction with the environment. Pushed by an initial prompt, its operation, however complex, rolls down a fixed (albeit probabilistic) and instantaneously reactive path toward the next output. And yet, the autoregressive aspect provides a powerful form of recursion that, within the right context, particularly that provided by "agency" extensions of LLMs, increases the interactional asymmetry property towards the transformer. When carefully crafted, LLMs can partially escape their downhill reactive nature (by rewriting the landscape they roll through) but, devoid of intrinsic individuation and normativity, fail to become genuine agents.

When measured against the criteria for autonomous agency derived from contemporary embodied and enactive frameworks, LLMs fall decisively short. This conclusion is grounded in the specific technical and operational details we outlined in Sect. 3. Their lack of self-sustaining individuality, their operation based on externally defined loss functions and reinforcements instead of intrinsic normativity, and their

²⁷ In a sense, it is possible to identify these systems as second and third order derivative intentional systems, following the classical distinction between *intrinsic* or genuine intentionality, characteristic of humans, and *derived* intentionality, characteristic of signs and artifacts (Haugeland, 1990; Searle, 1983). LLMs get their derived intentionality from that of their training data (second-order derived intentional systems) and, in turn, generate intentionally-structured products that recursively shape (through chain-of-thought techniques) their subsequent outputs (third-order derived intentionality). A more detailed account of this connection between Generative AI and intentionality can be found in Barandiaran and Pérez-Verdugo (2025).

fundamentally reactive nature are core architectural and operational limitations that preclude genuine agency.

Once established what LLMs are not, we are now faced with the constructive task of defining what they are. The next section, therefore, moves beyond critique to propose a positive conceptual framework for understanding their unique mode of existence and their transformative potential for human agency.

5 What are transformers then, if not agents, and how do they transform agency?

5.1 The language automaton and the ghost in the human-machine interaction

If transformer architectures, and LLMs as we know them, are not (autonomous) agents... What are LLMs that have such a strong impact on our digital environment and the way we live in them? If we are not to embrace transformers as members of our familiar way of being (as agents) in the world, we need to start somewhere else. First, it might be useful to distinguish operations from actions. An *operation* is a sequence of occurrences that can be interpreted functionally; that is, in (instrumental means-ends) relation to a final end state. Mechanisms carry out operations. A *digital operation* is a logical transformation executed by a computer. Operations are externally defined, whereas actions are internally defined by the agent that carries them out (in the sense outlined in the previous section). Note that an action can externally be defined as an operation and translated into a machine. But it is more convenient to use a specific name to label those processes that can be described both as an operation when carried out by a machine and as an action when done by a human: we might call those *tasks* (see Table 1).

With these distinctions at hand, we can now proceed to properly characterize LLMs. From the point of view of their organization, transformers are *automata*, as opposed to autonomous systems. This distinction is still relevant and crucial. Automatic²⁸ systems do not need human intervention to carry out certain operations, but are not autonomous. They cannot define their own identity and norms. However, as structured and identifiable digital instruction sets in physical memory and processors, they can carry out complex sequences of operations in the real world. They can transform energy into operations without human supervision or intervention during the process. They are *automata*.

However, transformers are not any kind of automata, they are of a very special digital kind, and operate in a very special type of environment, with an even more singular relationship to human life: they are *digital language automata* operating in multiple language-supporting and language-driven digital networks we continuously inhabit as linguistic animals (together with many other digital and physical objects around us).

²⁸ Although the etymology of automatic or automata ultimately brings us to self-minded or self-willed with a very strong mentalistic connotations, its popularization to depict mechanisms capable of completing sophisticated chains of operations is the sense that it nowadays embodies.

Table 1 Conceptual divide between machine and human types of identities, properties, and types of behavior. The middle column captures a common vocabulary that permits both humans and machines to share an interactive conceptual space. Note that systems can combine machine-individual hybrids to become individuals and individuals are (also) systems and can perfectly be interpretable as performing automatically

LLM	either	Human
Machine	System	Individual
Operation	Performance/Task	Action
Automata	Performer	Agent
Automatic	Performant	Autonomous
End state	Goal	Purpose
	Calculator	
	Writer	
	Painter	
	Interlocutor	

Moreover, LLM-powered chatbots, like ChatGPT, are specifically constituted by the way they couple with their associated milieu: other humans. As such, they often become *phantasmic language automata*²⁹. In some sense, ChatGPT is certainly the *ghost* of the text corpus that it originally was trained on. As a “phantom”, GPT can talk to us without the capacity to bring itself to life. Nevertheless, it is capable of conjuring all the knowledge of the corpus, of which itself is a shadow. As an automaton, it is capable of producing changes as a result of a sophisticated mechanism. Changes that trigger the response of the acknowledgment of an equal from those with which it phantasmatically interacts. In this sense, it is possible to characterize ChatGPT as an *interlocutor automaton* in a double sense that mirrors the twofold meaning of the term “interlocutor”: (a) as a system capable of performatively interacting in effective conversations with humans (and with other machines traditionally designed to take human produced text as input, e.g., programming, database queries, etc.); and (b) as an intermediary between (practically all digitized textual) human knowledge and other humans or machines.

In this sense, ChatGPT operates as a gigantic *text-that-talks*, or rather a *library-that-talks*³⁰, enabling a dialogical engagement with the vast corpus of human knowl-

²⁹ Our use of metaphors such as ‘ghost’ and ‘phantom’ is a deliberate methodological choice. In periods of rapid technological change, where the conceptual landscape is unsettled, metaphors serve as essential cognitive and philosophical tools for structuring an inquiry into a novel phenomenon (Hesse, 1966; Lakoff & Johnson, 1980). They function by inviting us to adopt a new perspective or “framing”, which organizes our understanding of a subject in ways that a literal description cannot (Camp, 2006). This allows us to take seriously the phenomenological character of the human-LLM interaction, which a purely technical description might obscure. Metaphorical phrasing allows us to condense intuitive understandings, anticipate future lines of inquiry, and reorganize our conceptual framework in ways that can precede and guide more detailed technical articulation. Our aim is to use these metaphors as conceptual probes, which we then ground in the technical and philosophical analysis that follows. In fact, the public and academic discourse surrounding LLMs is already deeply shaped by metaphors, from the deflationary “stochastic parrots” to “autocomplete on steroids”. Our goal is to provide new metaphorical framings that we believe more adequately capture the complexity and transformative potential of these technologies. In this case, the metaphor of a “phantasmatic language automaton” captures a dual nature. It is an automaton because its operations are mechanistically determined and devoid of the self-generated norms characteristic of genuine agency. The ‘phantasmic’ aspect, however, arises not from the machine alone, but from the human user’s phenomenological co-creation of the interaction. Humans readily project and enact a conversational partner when presented with a sufficiently coherent interlocutor. The LLM serves as a perfect scaffold for this projection, creating a ‘phantom’ conversational-other that is experientially real for the user but lacks an autonomous source on the machine’s side.

³⁰ And, we could say, in the more advanced multimodal cases, as a media-library-that-talks-and-paints.

edge and cultural heritage it has “internalized” (compressed on its transformer multidimensional spaces) and that it is capable of recruiting effectively in linguistic exchange. The machine’s interlocution, though devoid of personal intentionality, bears the trace of human experience as transposed into digitized textuality. The purpose-structured and bounded automatic interlocution, however, can be experienced as a genuine dialogue by the human subject. As a result, ChatGPT brings all this digitized textual knowledge, within the technological milieu, *ready-to-chat*.

Readiness to chat also implies *readiness-to-command*. And since we live in a world of linguistic performance and greatly digitized linguistic milieu, commanding can easily be turned into performance by a linguistic automaton. This is where recent enthusiasm with LLM “autonomous agents” lies. Transformers (properly “embodied”) can command themselves and other LLMs. Unlike mechanical automata, linguistic automata can be commanded by goals expressed in natural language, and can, at least partially, evaluate whether the results of the tasks carried out match the linguistic goal-expression. The capacity that truly brings linguistic automata close to some aspects of human agency is that commands of this sort need not be expressed within the strict boundaries of a computationally interpreted language (like programming languages or shell commands). LLMs can operate within the flexible and context dependent mesh of “natural” language.

Yet, the phantasmic dimension of LLM chatbots does not only reveal itself in its capacity to bring the dead text into non-living automatisms. It also has to do with ChatGPT as an enacted other, when given the interlocuting role. To appreciate the complexity of our interaction with ChatGPT, it is first essential that we understand what happens when we speak with other autonomous linguistic agents and what happens when we speak with “phantoms”. Embodied and enactive approaches to sociality (Di Paolo et al., 2018; Gallagher, 2017; Pérez & Gomila, 2021) defend that social cognition involves dynamic interaction with others. Social cognition is not about rationally reconstructing the thoughts that others are holding, but the result of ongoing fluent interactions between two or more agents. It has fundamentally more to do with dancing with another person than rationally strategizing a chess play by mail. Conversations are constituted by a complex chain of partial acts that imply, and somewhat anticipate, their completion by others (e.g., giving and taking, question and answer, salutation and response, etc.). What happens when the other is absent? Well, in a sense, we often play as-if it was there. We imagine that it is there, and re-enact a completion of our acts (e.g., an imaginary conversation with a friend). In a sense, we incorporate the absent other, internalizing the various dimensions of what would otherwise be an open interaction.

The experience of a phantom is somewhat a continuation of this capacity, to which we add the perceptual (visual, sound, etc.) hallucination as a means of a partial externalization of the experience. With ChatGPT, we have *excorporated* the phantom. The phantom is “real”, but still a phantom. The perceptual feedback is real (the text and sounds we hear come really from out there) but some hallucinatory and self-compleitive aspects of the interaction are still constitutive of it. There is no real-agent on the other side, but we still act as-if there was, thus in a sense making it a real conversational experience (see also Persson, 2024). Similar phenomena have been reported in Human-Robot Interaction research, describing the human tendency to navigate a

‘social artifact puzzle’ by treating mechanical systems as intentional agents to make interaction more tractable (Ziemke, 2023).

5.2 Transformer embodiments

Cognitive science has turned from abstract symbolic computations into the (historically neglected) role of the material body in the production of mindful experience and capacities (Calvo & Gomila, 2008; Gallagher, 2023; Shapiro, 2019; Varela et al., 1991). Cognitive agency is said to be embodied, extended beyond the brain as the “mere” hardware of the mind executing the genuine mindful “immaterial” software, into the living body of the cognizer, its technical equipment and sociomaterial environment. There are many layers of embodiment that Transformers rest on, and many that they lack compared to those of human intelligence. Most notably, ChatGPT lacks a sensorimotor body that captures variations on its physical environment (it lacks physical sensors) and also motor actuators that change its relationship with the environment and induce further sensory changes (Chemero, 2023)³¹. But, as we mentioned in the introduction, a sensorimotor robotic body is quickly being integrated with existing multi-modal LLMs, and its impact on deeper linguistic and behavioral skills can be expected to be significant. However, so far, transformers lack a living and lived body (real or simulated), that is often associated with emotions and affectivity regarding cognitive or agentive capacities (Damasio, 1994; Thompson, 2010). This certainly sets humans apart from LLMs. But there is nothing new or specific to LLMs on such lack of embodiment. In the previous section, we have sketched some consequences of this not-being-alive. This is a notable difference between Artificial and Natural intelligence that has attracted attention and arguments elsewhere as well (Froese & Ziemke, 2009; Koch, 2019; Seth, 2021; Thompson, 2010; Ziemke & Lowe, 2009). We shall now focus instead on two forms or aspects of LLM embodiment that are novel and have received little attention from the point of view of agency.

As a first type of embodiment (although in a sense very different from the usual one), we should focus on the *written corpus* (body in Latin) on which LLMs are trained and that they so effectively bring back to text. This is not simply a metaphorical use of the term body or embodiment. Textuality is an abstract, yet very concrete and complex, form of materiality itself, an externalized embodied product of linguistic agency. It brings with it purpose- and experience-structured relationships that might bear deep isomorphism with them. As human digitized culture is partly accessible to us, we often neglect the tremendous value (and size) of LLMs compressed *corpus*: an organized model of the textualized human knowledge and culture³². We

³¹ Modern multimodal systems identify and correlate patterns across modalities but lack the active, world-directed and normatively endowed nature of embodied agency. The crucial feedback loop, where an agent’s actions generate new sensory contingencies that shape subsequent actions, remains absent. The model analyses a static image, but it does not visually and actively explore a normatively valued environment to achieve an intrinsic goal. Thus, multimodality expands the model’s input beyond pure text, but it does not provide a sensorimotor body in the enactive sense. It merely expands the “library” of passively acquired data into a “media-library”.

³² We are so immersed in the complexities and subtleties of our culture and so “shocked” by the power of the new AI that we frequently only focus on what still distinguishes us from it or assign purely instrumental

need a theory of what this written corpus really is from an embodied and phenomenological cognitive science perspective. But current theories, so far, lack an account of the cognitive or agentive implications of large-scale computable textuality; and of the transformation of social, cognitive and ultimately biological lifeforms they bring with it.

For some theorists, it could well have been the case that a new generation of AI-systems was bootstrapping itself to human level intelligence by means of pure rational deduction and interaction with the environment (perhaps also as the exclusive result of the engineering effort of a private corporation³³). That is not the AI we have available today. This one is built on the collaborative effort of thousands of mathematical and computer science contributions, it is fed or trained on huge amounts of universal written heritage, collaborative digital commons (like Wikipedia or massive open source code repositories) and millions of distributed conversations and cooperative efforts (mostly on internet forums). No less important are the *embodiments of care* that usually take the form of labor externalization of massive data curation and operational alignment supervision, and that requires and recruits social-emotional resources from underpaid labor to safe-rail the brute models (Perrigo, 2023; C. Xiang, 2023). Another significant aspect of LLMs embodiment is their heavy computational and thus energetic and resource-hungry nature, and the extreme capitalist supply-chain extractivism it triggers, demands, and sustains (Valdivia, 2024).

In this sense, more than a self-bootstrapped Artificial Intelligence, ChatGPT, as an interlocutor automaton, is a computational proxy of the human collective intelligence externalized into a digitalized written body. It is, in turn, shaped and taken care of by hundreds of human and non-human lives. Thus, although in a very wide sense, yet one that is crucial to the effective operation of LLMs, transformers are embedded on large scale human and ecological bodies. This happens not just at a contextual level or as an operational environment, but at a constitutive level. No LLM is an island. And their performative power, and derived agentive capacities (if any), inherently rest on human and planetary scale life.

These dimensions of the material embodiment of LLM training and execution are crucial to understand the types of asymmetries it will bring to extended agentive capacities in humans. During the last 40 years, with the advent of the PC, there was little extended-agency computational asymmetries between human agents in general. Access to specific types of information has always been asymmetric between humans, but beyond this, the informatic capacity of a 14-year-old hacker and that of a big corporation was relatively even. With some rare exceptions, you and I could do the same thing with a PC or a mobile device compared to what Elon Musk, Queen Elizabeth or Warren Buffett could do. The computing (energy, processing, and data) cost of compiling and executing the most complex of software products (e.g., an operating system) was relatively accessible to most. Now, although (relatively low cost) access

value to it. However, how would we qualify the value of a LLM-like device, were it the only or primary source of access to an extinct human culture (or a distant alien one) for which we (or the LLM) possess (nevertheless) some kind of translation capacity?

³³ Ideally, also, that corporation produced or acquired the information and environment for the robot to become smart and also assumed the cost of externalities associated with its AI's intellectual growth.

to the best generic LLMs is still “widely” guaranteed, and despite locally-executable LLMs’ quality is growing, the asymmetry on the capacity to train and deploy specific LLMs, is orders of magnitude bigger.

On the one hand, there is the power implicit on how and what to train LLMs on, a power in the hands of those very few with the resources to train and shape a foundation model: the direct constraining power of training data choices and training procedure selection, the power of establishing RLHF criteria, constitutional writing capacity and interface design of massively deployed LLMs. On the other hand, the so-called “AI autonomous agents” only operate with some efficiency under exponential computational costs, either by a “social” distribution of tasks, by massive parallel planning and evaluation or by deeply iterative and long “chains of thought”. That is, by making up for the lack of genuine intrinsic purposefulness with redundancy. Their effective deployment might bring super-agency to privileged human masters, while delivering low-quality, low-cost alternatives to most. The asymmetry that the subsequent power amplifies is enormous. This is certainly not unlike the recursive power asymmetries that are already present in contemporary societies³⁴ except for a crucial fact: language digital automata remove humans (and their capacity for disobedience, resistance, and uprising) from the (social) sources of power.

5.3 Assisted, extended and midtended agency

What is ChatGPT as an extension of human cognitive agency? Is it an *assistant*? Most technologies that have been thought of as extensions or embodiments of human activity have been conceived as bodily prostheses. The extended mind hypothesis (Clark & Chalmers, 1998) and its later developments, including material engagement theories (Malafouris, 2016), have focused on the way in which beyond-the-skull extended material or computational processes should be understood as constitutive of cognition or brain processes. According to this view, mindful thinking processes must often be understood as extended into the material environment that they shape and, in turn, also shapes cognitive processes. Thinking involves *thinging*. When we make pottery, we do not print or carve an internal mental 3D model into the clay, we mold it. The materiality of the clay guides us, we bring the jar into form through a continuous reciprocal interaction between brain, body and world. Digital technologies constitute a branch of these phenomena: we offload memories, drafting procedures, image manipulation, etc. into our PC and mobile devices. Cognitive gadgets are organs of our mind extended beyond the skull.

Social interaction also extends and assists human agency: crew and team members interact (coordinatively or subordinatively), to achieve levels of agency (collective or directive) that are unreachable to a single individual (Lewis-Martin, 2022). The interlocutory capacity of LLMs brings human autonomous agency to a level not-unlike that of intersubjectively augmented agency, where the machine takes the role of an assistant (or that of a master). Recent human-computer interfaces have been dominated by action directed design. We tell computers what they have to do (by program-

³⁴ The richer you are, the higher your capacity to hire good lawyers to reduce your tax payments, your capacity to pay consultants to make more profit from your investment, etc.

ming a specific function, by pushing a button, by dragging a file, by selecting a menu item). LLMs, instead, make it possible to prompt or command an intention (Nielsen, 2023): “I want a summary of this text in French so that the 5-year-old child of my visiting friend can understand it”, “I need an impressionist style picture for a book cover on philosophy of mind”. This is certainly going to bring us to unprecedented forms of *master-slave* dialectical relationships with machines (and the corporations they ultimately serve). On the one hand, we might soon have at our disposal an “unlimited” number of assistant automata ready to perform computational and linguistic tasks for us. On the other hand, we might be equally commanded by them³⁵, not only through textual interfaces but also more sensorimotorly embodied throughout the course of our everyday behavior (e.g., Meta glasses connected to LLMs capable of interpreting our visual scenes and delivering “suggestion” or “orders” to accomplish specific tasks (Meta, 2023; Waisberg et al., 2023)).

But Transformers are also bringing with them a much deeper meaning of extended agency (with deeper dialectical connotations). There is a form of extended agency that LLMs already offer that becomes more intentionally intimate than any previous known form. In fact, this *extensional* character is closer to the *intentional* character of the mind that deserves a proper name: *midtentional*³⁶. We might best illustrate this form of mixed enhanced agency with some type of LLM integration on programming and office environments. But let’s first stop to analyze the phenomenology of non-AI-assisted writing. The process of writing (in paper or on the screen) is one that is often experienced as the very act of writing driving itself the intentions of the writer: the interaction process of writing pulls agency out of the head. It is the recurrent hand-keyboard-PC-screen-vision-brain-body-hand loop that produces text. Yet, we do not only write. We also supervise and edit recurrently. Thus, at least two loops are involved here, one is more pulled by the direct writer-text dynamics, the other by the more detached editorial supervision that either continuously or intermittently follows the former. At times, one finds the non-consciously written text as right and owned, as proper, and it is left untouched. Other times... “That is not what I meant exactly” ensues, “it needs a rewrite”. Both loops are person anchored. The environment (pen and paper or keyboard and screen) served as a support structure, a well-integrated, creative scaffold, providing the material basis of extended memory, recomposition, and tinkering. But the writer was the extended agent, the organized center of the scaffolded subject. This might start to change.

The enormous complexity and regulatory capacity of the brain-body system (compared to that of the passive materiality of the tool and work environment) is now chal-

³⁵ A step further on the already widespread tendency to be systematically commanded by apps running on real time data and AI driven optimization of work (like Uber, Glovo, etc.).

³⁶ The ‘mid-’ prefix is chosen to situate this process in the middle ground between purely *internal* sourcing of intentionality, and classical *extended* intentionality, where cognition is offloaded onto passive environmental resources. *Midtention* describes the hybrid process that emerges when an external system, itself possessing generative capacities (like an LLM, or another human), actively and reciprocally intervenes in the agent’s ongoing stream of active externalized thinking (like writing, designing, composing, coding, etc.) resulting on products that the autonomous agent takes as her own; yet the intentional process has been somehow displaced to an inter-mediate(d) space between the agent and a generative assistant. For a detailed account of this concept of “midtended cognition” see Barandiaran and Pérez-Verdugo (2025).

lenged by an ongoing activity of language automata, which are constantly reading us and writing (for) us. The extended autocomplete experience that LLMs provide can be tuned to integrate previous documents and styles of the writer, it mobilizes background knowledge, and is context-sensitive and purpose-structured (almost as if your shadow could push you into the direction you are intending to move towards). By feeding the LLM with the input tokens of the collaboration between its past output tokens and those written by you, the autocomplete feature adapts as you type. It often provides a mixed sense of agency, where what you find pre-written in the screen is “yours” without it having been necessarily written by you³⁷.

This brings the power of transformer-human interaction closer to a proper *cyborg agency*, beyond any experience of instrumental, social or intersubjective agency we might have ever encountered before (for a detailed account of cyborg intentionality, albeit pre-GPT, see Verbeek, 2008). In a textualized manner, this form of autocomplete is equivalent to injecting predictive efferent signals into the body movement. It is a step far beyond the classical examples of extended mind, in which, despite an out-of-the-skull spread of cognitive processes (the notebook, the mobile memory storage, the pen-and-paper calculation), the complexity asymmetry and integrative capacity was always tilted towards the skull-side of the coupled agent-environment system. It implies a degree of intimate technical transparency (Andrada et al., 2023; Pérez-Verdugo, 2022) in generative activities that challenges the very nature of human agency.

If we take predictive processing theories at face value (Clark, 2013, 2023; Friston, 2009; Seth, 2014, 2021) we might be encountering, for the first time, that the environment is delivering to the brain-body the very predictions that the brain-body is about to make about the effect of its own activity on the environment. It is tempting to consider this as a short-circuiting of agency as we know it. And it is difficult to assess the full consequences of the improvement of this technology, its massive adoption and multimodal expansion (introducing, perhaps, a new chapter on the many ethical implications of cognitive technologies, see Clowes et al., 2024). This brings the interlocutory nature of LLMs to an *intra*locutory mode of existence that uncannily blends with us.

6 Discussion and conclusions

The irruption of LLMs on our digital world, and through it on our lives (digital or otherwise), is breaking (again) what we thought human artifacts could never do. They perform operations that would require high levels of complex, common-sense, unstructured, and creative intelligence if performed by humans. We are forced to question their ontological status and the deep way in which they can transform ours. In this paper, we have focused on responding to this question, focusing on agency.

We started identifying the polarized take on LLMs ontological status: from their inflationary characterization as fully “sentient beings” to the deflationary one as mere “stochastic parrots”. Next, we delved into the faulty yet outstanding capacities

³⁷ In computer coding, this brings the expression of “predictive coding” to a completely new level.

that LLMs display, as measured by different human-level intelligence benchmarks. We concluded that a deeper delving into their organization is necessary to properly determine their mode of being. A detailed explanation of the complex and powerful architecture and processing of LLMs was provided next, together with the training and tuning procedures used to shape them. We also explained the different techniques that are used to enhance these systems to achieve agentic capacities. Turning to contemporary embodied philosophy of mind made it possible to identify what is missing from current models for them to achieve the status of genuine autonomous agents. We next moved to make a positive proposal as to how to treat LLMs and the way they transform our agency. It is time now to wrap up some concrete conclusions out of this journey.

6.1 Autonomous agents, interlocutory automata, and deeply embodied midtention

You are enmeshed in a thick web of recurrent attention-intention loops, of which you are both cause and effect. Through the growth and arrangement of this web, you have developed a sensitivity to navigate and stir your behavioral world so as to care for its deep precariousness. Along this open process, what is continuously guiding your action is not the anticipation of the next token of a pre-given text (or data-stream), but the sensitivity to the consequences of your actions at different nested scales: on the task at hand, on your goals, plans and, ultimately, on your own identity (itself the result of your own actively sustained and stirred encounters within the world). You are a genuine autonomous and organic agent. LLMs are not. But they perform, historically-unprecedented, extraordinary tasks. And they will continuously intersect with the way in which we navigate our (linguistic) worlds.

If by autonomous agent we mean an automatic system that is efficient on a sequence of multiple tasks, then LLMs (with important extensions and within structured digital environments) might already deserve the name. If, by agency, we mean the sense of agency we experience as autonomous self-defining and self-governing systems, then there are good reasons to believe that the LLM architecture as we know it currently falls short to meet the demands. This might not be a bug but an intended safety feature of transformer architectures. Systems that display complex intentional capacities might be a powerful assistant at a high price. Autonomous agents of this kind are the most powerful and yet most dangerous assistants. Being capable of creating your own norms, and being adaptively capable of displaying complex strategies to meet and transform them, is compatible with accepting external commands and making external norms your own. But it is also open to revolt. And this, in turn, opens a whole set of problems of AI alignment and safety.

A fundamental question remains: is it possible to achieve efficient and automatic multi step task-completion without genuine agency? Some strong requirements (like deep material living embodiment) might never be met by transformer-like systems. But it is still possible to envision variations on the current architecture that could bring the system's operations closer to living actions.

The deep transformation we are witnessing bears some relevant parallels with the industrial revolution and the increasing factory automation. Different degrees

of automation work well within factories, thanks to the operational constraints provided by the assembly line (always under human supervision and care). This time, the internet is the assembly line (rather network) for linguistic production. A LLM, put into the right sociotechnical environment (its “associated milieu”, Simondon, 2017), becomes an *interlocutor*, a task *performer*, or even an *operator* of the language fabric and the visuomotor internet (accessible through a web browser and keyboard-mouse actuators). The machine’s interlocutory capacity is real when coupled to other systems. It is certainly devoid of personal intentionality, but it brings humans to conversational life (as a clinical language automaton capable of supporting life without being itself alive). It is capable of this by means of the complex (context-sensitive, non-linear and massively parallel) re-generation of the multiple traces of human experience that the corpus of digitalized textuality embodies. Devoid of purposefulness, the intensive data training and human-machine guided reinforcement procedures make LLMs, however, purpose-structured and purpose-bounded systems that can command and be commanded.

When intimately coupled with human digital activity (and its history), LLMs can augment existing forms of agency in various ways. Some are well captured by the concept of “assistant” and involve forms of agency enhancement similar to those achieved by social coordination or subordination. Other forms of extending human agency are new. They involve the intervention of LLMs on the very ongoing activity of the human agent, by anticipating (based on context and previous history) the next token action(s) recursively. We called this *midtended* agency, in which machine operations blend with the subject into a unifying intentional process.

In order to understand the new technological landscape that LLMs open, we do not need (yet) to sacrifice the distinctive character of our autonomous agentic capacities. But we need to gain a detailed understanding of the capacities and dialectical processes that such systems will trigger. Language automata are here to stay, and we need to tune our conceptual systems to accommodate them.

6.2 On the dangers of the “stochastic parrot” metaphor and the “agency without intelligence” conceptualization

Deflationary accounts of AI tend to forget that human agency can, at a very fundamental level of quantum mechanics or the less fine-grained level of neuronal modeling, be characterized simply as a collection of dumb “probabilistic” and “stochastic” processes. It is not the description of the basic mechanisms that compose a complex system what defines its properties, but the organization of interrelated processes (both internal and interactive) that such basic mechanisms make possible to emerge. This is as true of us (living humans) as it is of any machine. In order to assess the genuine capacities of a system, we need to look at their internal workings, and the emergent capacities they can display when organized and coupled in specific manners. We need careful conceptual crafting to approach systems in which the relationship between the description of local mechanisms and the display of interactive capacities spans so many orders of magnitude (around 20) that we can hardly grasp.

LLMs are nothing alike stochastic parrots, nor domesticated living animals nor stochastic engines, neither caged or free in the rainforest, but effectively coupled

to the digital fabric of our social life. Whereas the metaphor of the “stochastic parrot” was once useful to question the rapidly emerging hype on LLMs (Bender et al., 2021), it might easily turn counterproductive. Parrots are living agents, ecologically balanced within their habitats, and capable of actively adapting to environmental changes (including those induced by human capture). They achieve this by means of deep cognitive, emotional and communicative capacities (far beyond the traditionally attributed dumb mimicry; see Pepperberg, 2006) that LLMs certainly lack. On the other hand, LLMs display capacities that effectively mobilize human intelligence as embodied in massive textuality, affectively mobilize human intelligence in conversation, and can activate forms of hybrid agency previously unavailable for human intelligence. And they do so by displaying powers far beyond those of stochastic, probabilistic or merely statistical token recombination.

Functionalist or informationalist conceptualizations do not play better than the “stochastic parrots” metaphor in the sociotechnical jungle. They fail to distinguish autonomous agency from mere digital processing. Declaring LLMs as “agents without intelligence” does not fix the foundational failure, it simply highlights it by unveiling the impotency to properly justify lack of intelligence on transformers. It also reverses the ontological order. They are better understood as (collective) intelligence without agency. The inversion does not only ignore the increasing material and energetic demands of AI and fuels LLM corporate marketing discourses. It misplaces our own agency and responsibility.

6.3 Prospects for transformed agencies in the era of deep digitality

Some of the properties that are essential to agency (most notably individuality and normativity) emerge from the deep materiality of organic agency. However, the recent course of AI explosion, with the gigantic investment of data and computational capacity (and the related energy demands) is revealing a *deep digitality* whose consequences are still to be fully unpacked. The complexity and scale of the operations involved in LLMs training and execution are huge. Prompt processing operations that, carried by an aware and conscious human, would take billions of years, challenge our intuitions and conceptual resources. By a digitality that deep, it is reasonable to hold that the boundary between invention and discovery, between artifact and nature, between engineering and science is somewhat blurred. We (humans) have built LLM as much as we have discovered their emergent capabilities³⁸. And avenues for a genuine digital agency might still be open for discovery. The way in which deep and wide materiality has revealed agentive capacities in natural history might well be somewhat replicated in the digital realm. If deep materiality brings with it the capacity to make difference emerge (Anderson, 1972) we have no reason to preclude the increasingly deep digitality of artificial devices to reveal new forms of agency, yet to come. And even deeper transformations of the existing forms of agency.

But depth alone does not bring matter (or digitality) to life. Ultimately, it is the organization of processes, their interaction with their environments, and their interdependence with the rest of beings, that needs to be scrutinized to disclose the mode of

³⁸ In fact, it is important to note how LLMs are rarely said to be built but “trained”.

existence of any device. No benchmark or general description (stochastic, statistical, probabilistic, syntactic, or otherwise) is sufficiently informative of the potential transformative capacities of machines. LLMs are no exception. Their mode of existence depends largely on human (and other) forms of life. And the deeper our materiality and digitality merge, the deeper will be the transformations to come. This is why transparency and openness regarding LLMs (and AI in general) is much more than a private ethical imperative and turns into a collective political concern: how these systems work and get coupled to our social fabric, how they feed on the human heritage and care, how they suck planetary resources, and affect social inequalities. To shape this future, we need a better conceptual understanding of how the mode of existence of LLMs transforms real agency.

Acknowledgements Special thanks to Steve Torrance, Miguel Aguilera, Enara García, Marta Pérez-Verdugo, Ekai Txapartegi, Ana Valdivia, Matthew Egbert, and Antonio Calleja-Lopez for careful and insightful revision and comments to early manuscripts of this paper. We are also grateful for the comments and revision of anonymous reviewers.

Author contributions A first draft of the comparison of LLMs with enactive theories of agency (Sect. 4) was written by Lola Almendros under the supervision of Xabier Barandiaran who then included the content of Sect. 3. Both together discussed the content of Sects. 5 and 6, drafted by Xabier Barandiaran and edited by Lola Almendros. Lola drafted Sect. 2 and Xabier Sect. 1 and both contributed to rewriting and reshaping content to its final form. Both authors read, carefully reviewed and approved the final manuscript.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. XEB and LSA acknowledge IAS-Research group funding IT1668-22 from Basque Government, grants PID2019-104576GB-I00 for project Outonomy, and PID2023-147251NB-I00 for project Outagencies funded by MCIU/AEI/<https://doi.org/10.13039/501100011033>. LSA also acknowledges the University of Salamanca's Margarita Salas Postdoctoral Grant Program within Spain's Recovery, Transformation, and Resilience Plan and Next Generation EU.

Data availability No relevant data or material was used for this paper other than the references provided.

Declarations

Ethical approval Does not apply.

Informed consent Does not apply.

Statement Regarding Research Involving Human Participants and/or Animals Does not apply.

Competing interests The author has no relevant financial or non-financial interests to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- admin_sagi (2023). *SuperAGI - Opensource AGI Infrastructure*. SuperAGI. <https://superagi.com/>
- Aguilera, M., Millidge, B., Tschantz, A., & Buckley, C. L. (2022). How particular is the physics of the free energy principle? *Physics of Life Reviews*, 40, 24–50. <https://doi.org/10.1016/j.plrev.2021.11.001>
- Anderson, P. W. (1972). More is different. *Science*, 177(4047), 393–396. <https://doi.org/10.1126/science.1177.4047.393>
- Andrada, G., Clowes, R. W., & Smart, P. R. (2023). Varieties of transparency: Exploring agency within AI systems. *AI & SOCIETY*, 38(4), 1321–1331. <https://doi.org/10.1007/s00146-021-01326-6>
- Andreas, J. (2022). *Language Models as Agent Models* (No. arXiv:2212.01681). arXiv. <https://doi.org/10.48550/arXiv.2212.01681>
- Anthropic (2025). *Claude 4 System Card*. <https://www.anthropic.com/claude-4-system-card>
- Anthropic (2024). *Introducing the Model Context Protocol*. <https://www.anthropic.com/news/model-context-protocol>
- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., & Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI Feedback* (No. arXiv:2212.08073). <https://doi.org/10.48550/arXiv.2212.08073>
- Barandiaran, X. E. (2007). Mental life: Conceptual models and synthetic methodologies for a post-cognitivist psychology. In B. Wallace, A. Ross, J. Davies, & T. Anderson (Eds.), *The world, the Mind and the body: Psychology after cognitivism* (pp. 49–90). Imprint Academic.
- Barandiaran, X. E. (2008). *Mental Life: A naturalized approach to the autonomy of cognitive agents*. [PhD Thesis, University of the Basque Country (UPV-EHU)]. <https://xabier.barandiaran.net/phdthesis/>
- Barandiaran, X. E., & Egbert, M. D. (2014). Norm-establishing and norm-following in autonomous agency. *Artificial Life*, 20(1), 5–28. https://doi.org/10.1162/ARTL_a_00094
- Barandiaran, X. E., & Moreno, A. (2006). On what makes certain dynamical systems cognitive: A minimally cognitive organization program. *Adaptive Behavior*, 14(2), 171–185. <https://doi.org/10.1177/105971230601400208>
- Barandiaran, X. E., & Pérez-Verdugo, M. (2025). Generative midtended cognition and artificial intelligence: Thinging with thinging things. *Synthese*, 205(4), 1–24. <https://doi.org/10.1007/s11229-025-04961-4>
- Barandiaran, X. E., & Rama, T. (2025). *Sensorimotor teleology and goal-directedness. An organismic framework for normative behaviour*. <https://philsci-archival.pitt.edu/id/eprint/25369>
- Barandiaran, X. E., Di Paolo, E., & Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and Spatio-temporality in action. *Adaptive Behavior*, 17(5), 367–386. <https://doi.org/10.1177/1059712309343819>
- Bayne, T., & Williams, I. (2023). The turing test is not a good benchmark for thought in LLMs. *Nature Human Behaviour*, 7(11). <https://doi.org/10.1038/s41562-023-01710-w>
- Bechtel, W., & Richardson, R. C. (2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. MIT Press.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Berg, H. C. (2004). *E. coli in motion*. Springer [u.a.].
- Bertolini, A., & Episcopo, F. (2022). Robots and AI as legal subjects?? Disentangling the ontological and functional perspective. *Frontiers in Robotics and AI*, 9, 842213. <https://doi.org/10.3389/frobt.2022.842213>
- Biever, C. (2023). ChatGPT broke the turing test—The race is on for new ways to assess AI. *Nature*, 619(7971), 686–689. <https://doi.org/10.1038/d41586-023-02361-7>
- Bostrom, N. (2017). *Superintelligence*. Dunod.
- Bousmalis, K., Vezzani, G., Rao, D., Devin, C., Lee, A. X., Bauza, M., Davchev, T., Zhou, Y., Gupta, A., Raju, A., Laurens, A., Fantacci, C., Dalibard, V., Zambelli, M., Martins, M., Pevcevičute, R., Blokzijl, M., Denil, M., Batchelor, N., & Heess, N. (2023). *RoboCat: A Self-Improving Generalist Agent for Robotic Manipulation* (No. arXiv:2306.11706). arXiv. <https://doi.org/10.48550/arXiv.2306.11706>
- Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choremanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., Florence, P., Fu, C., Arenas, M. G., Gopalakrishnan, K., Han, K., Hausman, K., Herzog, A., Hsu, J., Ichter, B., & Zitkovich, B. (2023). *RT-2: Vision-Language-Action models transfer web knowledge to robotic control*.

- Brooks, R. A. (1990). Elephants don't play chess. *Robotics and Autonomous Systems*, 6(1–2), 3–15.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., & Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901. <https://papers.nips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>
- Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know?? Practical estimates of vocabulary size dependent on word definition, the degree of Language input and the participant's age. *Frontiers in Psychology*, 7. <https://doi.org/10.3389/fpsyg.2016.01116>
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4* (No. arXiv:2303.12712). arXiv. <https://doi.org/10.48550/arXiv.2303.12712>
- Burnell, R., Schellaert, W., Burden, J., Ullman, T. D., Martinez-Plumed, F., Tenenbaum, J. B., Rutar, D., Cheke, L. G., Sohl-Dickstein, J., Mitchell, M., Kiela, D., Shanahan, M., Voorhees, E. M., Cohn, A. G., Leibo, J. Z., & Hernandez-Orallo, J. (2023). Rethink reporting of evaluation results in AI. *Science*, 380(6641), 136–138. <https://doi.org/10.1126/science.adf6369>
- Butkus, M. A. (2020). The human side of artificial intelligence. *Science and Engineering Ethics*, 26(5), 2427–2437. <https://doi.org/10.1007/s11948-020-00239-9>
- Buzsaki, G. (2006). *Rhythms of the brain* (1st ed.). Oxford University Press.
- Calvo, P., & Gomila, T. (2008). *Handbook of cognitive science: An embodied approach*. Elsevier Science.
- Camp, E. (2006). Metaphor and that certain Je Ne Sais quoi. *Philosophical Studies*, 129(1), 1–25. <https://doi.org/10.1007/s11098-005-3019-5>
- Carlsmith, J. (2020). *How Much Computational Power Does It Take to Match the Human Brain?* [Research Report]. Open Philanthropy. <https://www.openphilanthropy.org/research/how-much-computational-power-does-it-take-to-match-the-human-brain/>
- Carruthers, P. (2006). *The architecture of the Mind*. Oxford University Press.
- Chebotar, Y., Vuong, Q., Irfan, A., Hausman, K., Xia, F., Lu, Y., Kumar, A., Yu, T., Herzog, A., Pertsch, K., Gopalakrishnan, K., Ibarz, J., Nachum, O., Sontakke, S., Salazar, G., Tran, H. T., Peralta, J., Tan, C., Manjunath, D., & Levine, S. (2023). *Q-Transformer: Scalable offline reinforcement learning via autoregressive Q-Functions*.
- Chemero, A. (2009). *Radical embodied cognitive science*. The MIT Press.
- Chemero, A. (2023). LLMs differ from human cognition because they are not embodied. *Nature Human Behaviour*, 7(11). <https://doi.org/10.1038/s41562-023-01723-5>
- Chen, J., Jiang, Y., Lu, J., & Zhang, L. (2024). S-Agents: Self-organizing agents in Open-ended environments. *ArXiv*. <https://doi.org/10.48550/ArXiv.2402.04578>. No. ArXiv:2402.04578.
- Chiang, T. (2023). ChatGPT Is a Blurry JPEG of the Web. *The New Yorker*. <https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web>
- Chollet, F., Knoop, M., Kamradt, G., Landers, B., & Pinkard, H. (2025). ARC-AGI-2: A new challenge for frontier AI reasoning systems. *ArXiv*. <https://doi.org/10.48550/ArXiv.2505.11831>. No. ArXiv:2505.11831.
- Chomsky, N., Roberts, I., & Watumull, J. (2023). The false promise of ChatGPT. *The New York Times*. <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>
- Christian, B. (2021). *The alignment problem: Machine learning and human values*. W. W. Norton & Company.
- Churchland, P. M. (1990). *On the nature of theories: A neurocomputational perspective*. <http://conservancy.umn.edu/handle/11299/185730>
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(03), 181–204. <https://doi.org/10.1017/S0140525X12000477>
- Clark, A. (2023). *The experience machine: How our Minds predict and shape reality*. Pantheon.
- Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7.
- Clowes, R. W., Smart, P. R., & Heersmink, R. (2024). The ethics of the extended mind: Mental privacy, manipulation and agency. In B. Beck, O. Friedrich, & J. Heinrichs (Eds.), *Neuroprosthetics: Ethics of applied situated cognition*. Springer.
- Coeckelbergh, M. (2021). Narrative responsibility and artificial intelligence: How AI challenges human responsibility and sense-making. *AI & SOCIETY*. <https://doi.org/10.1007/s00146-021-01375-x>

- Collaboration, O. X. E., Padalkar, A., Pooley, A., Jain, A., Bewley, A., Herzog, A., Irpan, A., Khazatsky, A., Rai, A., Singh, A., Brohan, A., Raffin, A., Wahid, A., Burgess-Limerick, B., Kim, B., Schölkopf, B., Ichter, B., Lu, C., Xu, C., & Cui, Z. J. (2023). *Open X-Embodiment: Robotic Learning Datasets and RT-X Models* (No. arXiv:2310.08864). arXiv. <https://doi.org/10.48550/arXiv.2310.08864>
- Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.
- Damasio, A. R. (1994). *Descartes' error*. G.P. Putnam.
- Davidson, D. (1980). *Essays on actions and events*. Oxford University Press.
- DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., & Zhang, Z. (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning* (No. arXiv:2501.12948). arXiv. <https://doi.org/10.48550/arXiv.2501.12948>
- Dennett, D. C. (1989). *The intentional stance*. MIT Press.
- Di Paolo, E. A., Buhrmann, T., & Barandiaran, X. E. (2017). *Sensorimotor life: An enactive proposal*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198786849.001.0001>
- Di Paolo, E. A., Cuffari, E. C., & De Jaegher, H. (2018). *Linguistic bodies: The continuity between life and language*. MIT press. <http://gen.lib.rus.ec/book/index.php?md5=9969BBD6CC722AE0EF427DE8C57585FC>
- Di Paolo, E. A., Thompson, E., & Beer, R. (2022). Laying down a forking path: Tensions between enaction and the free energy principle. *Philosophy and the Mind Sciences*, 3. <https://doi.org/10.33735/pnimisci.2022.9187>
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houshy, N. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* (No. arXiv:2010.11929). arXiv. <https://doi.org/10.48550/arXiv.2010.11929>
- Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. The MIT Press.
- Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., Welleck, S., West, P., Bhagavatula, C., Bras, R. L., Hwang, J. D., Sanyal, S., Ren, X., Ettinger, A., Harchaoui, Z., & Choi, Y. (2023). *Faith and Fate: Limits of Transformers on Compositionality*. Thirty-seventh Conference on Neural Information Processing Systems. <https://openreview.net/forum?id=Fkckkr3ya8>
- Favela, L. H. (2020). Dynamical systems theory in cognitive science and neuroscience. *Philosophy Compass*, 15(8), e12695. <https://doi.org/10.1111/phc3.12695>
- Floridi, L. (2007). A look into the future impact of ICT on our lives. *The Information Society*, 23(1), 59–64. <https://doi.org/10.1080/01972240601059094>
- Floridi, L. (2023). AI as agency without intelligence: On chatgpt, large Language models, and other generative models. *Philosophy & Technology*, 36(1), 15. <https://doi.org/10.1007/s13347-023-00621-y>
- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4), 681–694. <https://doi.org/10.1007/s11023-020-09548-1>
- Fourneret, E., & Yvert, B. (2020). Digital normativity: A challenge for human subjectivation. *Frontiers in Artificial Intelligence*, 3, 27. <https://doi.org/10.3389/frai.2020.00027>
- Frankfurt, H. G. (2005). *On bullshit*. Princeton University Press.
- Freeman, W. J. (2001). *How brains make up their Minds* (1st ed.). Columbia University.
- Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, 13(7), 293–301. <https://doi.org/10.1016/j.tics.2009.04.005>
- Froese, T., & Ziemke, T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence*, 173(3), 466–500.
- Gallagher, S. (2017). *Enactivist interventions: Rethinking the mind* (First edition). Oxford University Press.
- Gallagher, S. (2023). *Embodied and enactive approaches to cognition* (1st ed.). Cambridge University Press. <https://doi.org/10.1017/9781009209793>
- Ge, W., Chen, S., Chen, G., Chen, J., Chen, Z., Yan, S., Zhu, C., Lin, Z., Xie, W., Wang, X., Gao, A., Zhang, Z., Li, J., Wan, X., & Wang, B. (2023). *MLLM-Bench, Evaluating Multi-modal LLMs using GPT-4V* (No. arXiv:2311.13951). arXiv. <https://doi.org/10.48550/arXiv.2311.13951>
- Gemini Team, Anil, R., Borgeaud, S., Wu, Y., Alayrac, J. B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Petrov, S., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., & Vinyals, O. (2023). *Gemini: A Family of Highly Capable Multimodal Models* (No. arXiv:2312.11805). arXiv. <https://doi.org/10.48550/arXiv.2312.11805>
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Routledge.

- Google DeepMind (2025). *Gemini 2.5 Pro: Model Card & Benchmark Results*. <https://deepmind.google/models/gemini/pro/>
- Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). *How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection* (No. arXiv:2301.07597). arXiv. <https://doi.org/10.48550/arXiv.2301.07597>
- Haugeland, J. (1990). The intentionality all-stars. *Philosophical Perspectives*, 4, 383–427.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). *Measuring Massive Multitask Language Understanding* (No. arXiv:2009.03300). arXiv. <https://doi.org/10.48550/arXiv.2009.03300>
- Heras-Escribano, M. (2019). *The Philosophy of Affordances* (1st ed. 2019 edition). Palgrave Macmillan.
- Hesse, M. B. (1966). *Models and analogies in science*. University of Notre Dame.
- Hessel, J., Marasović, A., Hwang, J. D., Lee, L., Da, J., Zellers, R., Mankoff, R., & Choi, Y. (2023). *Do Androids Laugh at Electric Sheep? Humor Understanding Benchmarks from The New Yorker Caption Contest* (No. arXiv:2209.06293). arXiv. <https://doi.org/10.48550/arXiv.2209.06293>
- Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology*, 26(2). <https://doi.org/10.1007/s10676-024-09775-5>
- Ho, J., Jain, A., & Abbeel, P. (2020). *Denoising Diffusion Probabilistic Models* (No. arXiv:2006.11239). arXiv. <https://doi.org/10.48550/arXiv.2006.11239>
- Holmes, A. (2024). To Unlock AI Spending, Microsoft, OpenAI And Google Prep Agents. *The Information*. <https://www.theinformation.com/articles/to-unlock-ai-spending-microsoft-openai-and-google-prep-agents>
- Hou, X., Zhao, Y., Wang, S., & Wang, H. (2025). *Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions* (No. arXiv:2503.23278). arXiv. <https://doi.org/10.48550/arXiv.2503.23278>
- Huang, X., Liu, W., Chen, X., Wang, X., Wang, H., Lian, D., Wang, Y., Tang, R., & Chen, E. (2024). *Understanding the planning of LLM agents: A survey* (No. arXiv:2402.02716). arXiv. <https://doi.org/10.48550/arXiv.2402.02716>
- Huben, R. (2023). How does GPT-3 spend its 175B parameters? [Substack newsletter]. *From AI to ZI*. <https://aizi.substack.com/p/how-does-gpt-3-spend-its-175b-parameters>
- Hutto, D. D., & Myin, E. (2012). *Radicalizing enactivism: Basic Minds without content*. MIT Press.
- Jebari, K., & Lundborg, J. (2021). Artificial superintelligence and its limits: Why alphazero cannot become a general agent. *AI & SOCIETY*, 36(3), 807–815. <https://doi.org/10.1007/s00146-020-01070-3>
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D., de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M. A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). *Mistral 7B*. arXiv. <https://arxiv.org/abs/2310.06825v1>
- Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D., de las, Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M. A., Stock, P., Subramanian, S., Yang, S., & Sayed, W. E. (2024a). *Mixtral of Experts* (No. arXiv:2401.04088). arXiv. <https://doi.org/10.48550/arXiv.2401.04088>
- Jiang, D., Zhang, J., Weller, O., Weir, N., Van Durme, B., & Khashabi, D. (2024b). *SELF-[IN]CORRECT: LLMs Struggle with Refining Self-Generated Responses* (No. arXiv:2404.04298). arXiv. <http://arxiv.org/abs/2404.04298>
- Johnson, M., & Lakoff, G. (2002). Why cognitive linguistics requires embodied realism. *Cognitive Linguistics*, 13(3). <https://doi.org/10.1515/cogl.2002.016>
- Jonas, H. (1966). *The Phenomenon of Life. Toward a Philosophy of Biology*. Chicago-London.
- Jones, C. R., & Bergen, B. K. (2024). *People cannot distinguish GPT-4 from a human in a Turing test* (No. arXiv:2405.08007). arXiv. <https://doi.org/10.48550/arXiv.2405.08007>
- Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., & Kaplan, J. (2022). *Language Models (Mostly) Know What They Know*. <https://doi.org/10.48550/ARXIV.2207.05221>
- Kambhampati, S. (2023). Can LLMs really reason and plan?— Communications of the ACM. *Communications of the ACM*. <https://cacm.acm.org/blogcacm/can-llms-really-reason-and-plan/>
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling Laws for Neural Language Models* (No. arXiv:2001.08361). arXiv. <https://doi.org/10.48550/arXiv.2001.08361>

- Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, 15(138), 20170792. <https://doi.org/10.1098/rsif.2017.0792>
- Knight, W. (2014). Forget Chatbots. AI Agents Are the Future. *Wired*. <https://www.wired.com/story/fast-forward-forget-chatbots-ai-agents-are-the-future/>
- Koch, C. (2019). *The Feeling of Life Itself: Why Consciousness Is Widespread but Can't Be Computed*.
- Kocijan, V., Davis, E., Lukasiewicz, T., Marcus, G., & Morgenstern, L. (2023). The defeat of the Winograd schema challenge. *Artificial Intelligence*, 325, 103971. <https://doi.org/10.1016/j.artint.2023.103971>
- Kubes, T., & Reinhardt, T. (2022). Techno-species in the becoming towards a relational ontology of Multi-species assemblages (ROMA). *NanoEthics*, 16(1), 95–105. <https://doi.org/10.1007/s11569-021-00401-y>
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by* (Vol. 111). Chicago London.
- Lawson, A. (2025). *Comment on The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity* (No. arXiv:2506.09250; Version 2). arXiv. <https://doi.org/10.48550/arXiv.2506.09250>
- Lemoine, B. (2022). Is LaMDA Sentient? — An Interview. *Medium*. <https://cajundiscordian.medium.com/is-lambda-sentient-an-interview-ea64d916d917>
- Levesque, H. J., Davis, E., & Morgenstern, L. (2012). *The Winograd schema challenge*. *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, 552–561.
- Lewis-Martin, J. (2022). What kinds of groups are group agents? *Synthese*, 200(4), 283. <https://doi.org/10.1007/s11229-022-03766-z>
- Li, H., Chong, Y. Q., Stepputtis, S., Campbell, J., Hughes, D., Lewis, M., & Sycara, K. (2023). *Theory of Mind for Multi-Agent Collaboration via Large Language Models*. 180–192. <https://doi.org/10.48550/arXiv.2310.10701>
- Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2023). *Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task* (No. arXiv:2210.13382). arXiv. <http://arxiv.org/abs/2210.13382>
- Li, J., Zhang, Q., Yu, Y., Fu, Q., & Ye, D. (2024). *More Agents Is All You Need* (No. arXiv:2402.05120). arXiv. <https://doi.org/10.48550/arXiv.2402.05120>
- Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., & Cobbe, K. (2023). *Let's Verify Step by Step* (No. arXiv:2305.20050). arXiv. <https://doi.org/10.48550/arXiv.2305.20050>
- Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., & Tang, J. (2023). AgentBench: Evaluating LLMs as Agents. *ArXiv, abs/2308.03688*. <https://doi.org/10.48550/arXiv.2308.03688>
- Mabaso, B. A. (2021). Computationally rational agents can be moral agents. *Ethics and Information Technology*, 23(2), 137–145. <https://doi.org/10.1007/s10676-020-09527-1>
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhunoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., & Clark, P. (2023a). *Self-Refine: Iterative Refinement with Self-Feedback* (No. arXiv:2303.17651). arXiv. <https://doi.org/10.48550/arXiv.2303.17651>
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhunoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., & Clark, P. (2023b). *Self-Refine: Iterative Refinement with Self-Feedback* (No. arXiv:2303.17651). arXiv. <https://doi.org/10.48550/arXiv.2303.17651>
- Malafouris, L. (2016). *How things shape the mind: A theory of material engagement*. MIT Press.
- Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition*. D. Reidel Publishing Company.
- Merleau-Ponty, M. (1944). *Phenomenology of perception*. Routledge.
- Meta, A. I. (2025). *The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation*. Meta AI. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>
- Meta (2023). Introducing the New Ray-Ban| Meta Smart Glasses. *Meta*. <https://about.fb.com/news/2023/09/new-ray-ban-meta-smart-glasses/>
- Mialon, G., Dessi, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., LeCun, Y., & Scialom, T. (2023). *Augmented Language Models: A Survey* (No. arXiv:2302.07842). arXiv. <https://doi.org/10.48550/arXiv.2302.07842>

- Miller, R. (2023). *Holding Large Language Models to Account*. <https://www.semanticscholar.org/paper/Holding-Large-Language-Models-to-Account-Miller/5360b929d1782d21ed71b5528fd546f9f15a4106>
- Mirchandani, S., Xia, F., Florence, P., Ichter, B., Driess, D., Arenas, M. G., Rao, K., Sadigh, D., & Zeng, A. (2023). *Large Language Models as General Pattern Machines*. <https://doi.org/10.48550/ARXIV.2307.04721>
- Mistral, A. I., Rastogi, A., Jiang, A. Q., Lo, A., Berrada, G., Lample, G., Rute, J., Barmentlo, J., Yadav, K., Khandelwal, K., Chandu, K. R., Blier, L., Saulnier, L., Dinot, M., Darrin, M., Gupta, N., Soletskyi, R., Vaze, S., Scao, T. L., & Tang, Y. (2025). *Magistral* (No. arXiv:2506.10910). arXiv. <https://doi.org/10.48550/arXiv.2506.10910>
- Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13), e2215907120. <https://doi.org/10.1073/pnas.2215907120>
- Moonshot (2025). *MoonshotAI/Kimi-K2* [Computer software]. Moonshot AI. <https://github.com/MoonshotAI/Kimi-K2> (Original work published 2025).
- Moreno, A., & Mossio, M. (2015). *Biological autonomy: A philosophical and theoretical enquiry*. Springer.
- Mueller, B. (2023). *MiniAGI* [Python]. <https://github.com/muellerberndt/mini-agi> (Original work published 2023).
- Nave, K. (2025). *A drive to survive: The free energy principle and the meaning of life (online draft)*. MIT Press. <https://osf.io/ds9mn>
- Newell, A. (1980). Physical symbol systems. *Cognitive Science*, 4(2), 135–183.
- Nielsen, J. (2023). AI: First new UI paradigm in 60 years. *Nielsen Norman Group*. <https://www.nngroup.com/articles/ai-paradigm/>
- Noë, A. (2004). *Action in perception*. The MIT Press.
- Nolan, C. (Director). (2001). *Memento* [Mystery, Thriller]. Newmarket Capital Group, Team Todd, I Remember Productions.
- Norvig, P., & Russell, S. (2021). *Artificial intelligence: A modern approach, global edition*. Pearson.
- O'Donnell, J. (2024). Sam Altman says helpful agents are poised to become AI's killer function. *MIT Technology Review*. <https://www.technologyreview.com/2024/05/01/1091979/sam-altman-says-helpful-agents-are-poised-to-become-ais-killer-function/>
- OpenAI (2024b). *Learning to reason with LLMs*. <https://openai.com/index/learning-to-reason-with-llms/>
- OpenAI (2025b). *ChatGPT agent System Card*. <https://openai.com/index/chatgpt-agent-system-card/>
- OpenAI (2024a). *GPT-4o System Card*. <https://openai.com/index/gpt-4o-system-card/>
- OpenAI (2025a). *OpenAI o3 and o4-mini System Card* [Technical report]. OpenAI. <https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf>
- OpenAI (2023). *GPT-4 Technical Report* (No. arXiv:2303.08774). arXiv. <https://doi.org/10.48550/arXiv.2303.08774>
- Pasquinelli, M. (2023). *The eye of the master: A social history of artificial intelligence*. Verso.
- Pepperberg, I. M. (2006). Cognitive and communicative abilities of grey parrots. *Applied Animal Behaviour Science*, 100(1), 77–86. <https://doi.org/10.1016/j.applanim.2006.04.005>
- Pérez, D. I., & Gomila, A. (2021). *Social Cognition and the Second Person in Human Interaction* (1st edition). Routledge.
- Pérez-Verdugo, M. (2022). Situating transparency: An extended cognition approach. *Teorema: Revista Internacional De Filosofía*, 41(3), 7–24.
- Perrigo, B. (2023). OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic. *Time*. <https://time.com/6247678/openai-chatgpt-kenya-workers/>
- Persson, A. (2024). *Cognition and the Machine* [PhD Thesis, Uppsala University]. <https://uu.diva-portal.org/smash/record.jsf?pid=diva2%3A1850218&dsid=389>
- Port, R. F., & Gelder, T. V. (1995). *Mind as motion: Explorations in the dynamics of cognition*. MIT Press.
- Putnam, H. (1965). The mental life of some machines. *Philosophical papers: Volume 2, mind, Language and reality* (pp. 408–428). Cambridge University Press.
- Qwen, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., & Qiu, Z. (2025). *Qwen2.5 Technical Report* (No. arXiv:2412.15115). arXiv. <https://doi.org/10.48550/arXiv.2412.15115>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*.

- Raja, V., Valluri, D., Baggs, E., Chemero, A., & Anderson, M. L. (2021). The Markov blanket trick: On the scope of the free energy principle and active inference. *Physics of Life Reviews*, 39, 49–72. <https://doi.org/10.1016/j.plrev.2021.09.001>
- Reed, E. S. (1996). *Encountering the world: Toward an ecological psychology*. Oxford University Press.
- Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-maroon, G., Giménez, M., Sulsky, Y., Kay, J., Springenberg, J. T., Eccles, T., Bruce, J., Razavi, A., Edwards, A., Heess, N., Chen, Y., Hadsell, R., Vinyals, O., Bordbar, M., & Freitas, N. (2022). A Generalist Agent. *Transactions on Machine Learning Research*. <https://openreview.net/forum?id=1ikK0kHjvj>.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536. <https://doi.org/10.1038/323533a0>
- Rumelhart, D. E., McClelland, J. L., & Group, P. R. (1987). *Parallel distributed processing, vol. 1: Foundations*. The MIT Press.
- Russell, S. J. (2019). *Human compatible: Artificial intelligence and the problem of control*.
- Sakaguchi, K., Bras, R. L., Bhagavatula, C., & Choi, Y. (2019). *WinoGrande: An Adversarial Winograd Schema Challenge at Scale* (No. arXiv:1907.10641). arXiv. <https://doi.org/10.48550/arXiv.1907.10641>
- Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). *Toolformer: Language Models Can Teach Themselves to Use Tools* (No. arXiv:2302.04761). arXiv. <https://doi.org/10.48550/arXiv.2302.04761>
- Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., & Silver, D. (2020). Mastering atari, go, chess and Shogi by planning with a learned model. *Nature*, 588, 604–609. <https://doi.org/10.1038/s41586-020-03051-4>
- Schwitzgebel, E., Schwitzgebel, D., & Strasser, A. (2024). Creating a large Language model of a philosopher. *Mind & Language*, 39(2), 237–259. <https://doi.org/10.1111/mila.12466>
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424.
- Searle, J. R. (1983). *Intentionality: An essay in the philosophy of Mind*. Cambridge University Press.
- Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience*, 5(2), 97–118. <https://doi.org/10.1080/17588928.2013.877880>
- Seth, A. K. (2021). *Being you: A new science of consciousness*. Faber & Faber.
- Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., & Guo, D. (2024). *DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models* (No. arXiv:2402.03300). arXiv. <https://doi.org/10.48550/arXiv.2402.03300>
- Shapiro, L. (2019). *Embodied Cognition*.
- Shardlow, M., & Przybyla, P. (2023). *Deanthropomorphising NLP: Can a Language Model Be Conscious?* (No. arXiv:2211.11483). arXiv. <https://doi.org/10.48550/arXiv.2211.11483>
- Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). *Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer* (No. arXiv:1701.06538). arXiv. <https://doi.org/10.48550/arXiv.1701.06538>
- Shen, M., & Yang, Q. (2025). *From Mind to Machine: The Rise of Manus AI as a Fully Autonomous Digital Agent* (No. arXiv:2505.02024). arXiv. <https://doi.org/10.48550/arXiv.2505.02024>
- Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., & Farajtabar, M. (2025). *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity* (No. arXiv:2506.06941). arXiv. <https://doi.org/10.48550/arXiv.2506.06941>
- Siegelmann, H. T., & Sontag, E. D. (1995). On the computational power of neural Nets. *Journal of Computer and System Sciences*, 50(1), 132–150. <https://doi.org/10.1006/jcss.1995.1013>
- Significant Gravitas (2023). *AutoGPT* [JavaScript]. <https://github.com/Significant-Gravitas/AutoGPT> (Original work published 2023).
- Simondon, G. (2017). *On the mode of existence of technical objects*. University of Minnesota Press.
- Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S. R., Cole-Lewis, H., Neal, D., Rashid, Q. M., Schaeckermann, M., Wang, A., Dash, D., Chen, J. H., Shah, N. H., Lachgar, S., Mansfield, P. A., & Natarajan, V. (2025). Toward expert-level medical question answering with large Language models. *Nature Medicine*, 31(3), 943–950. <https://doi.org/10.1038/s41591-024-03423-7>
- Sprague, Z., Ye, X., Bostrom, K., Chaudhuri, S., & Durrett, G. (2023). *MuSR: Testing the Limits of Chain-of-thought with Multistep Soft Reasoning* (No. arXiv:2310.16049). arXiv. <https://doi.org/10.48550/arXiv.2310.16049>

- Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., & Wu, Z. (2023). *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models* (No. arXiv:2206.04615). arXiv. <https://doi.org/10.48550/arXiv.2206.04615>
- Sterelny, K. (2001). *The evolution of agency and other essays*. Cambridge University Press.
- Thompson, E. (2010). *Mind in life: Biology, phenomenology, and the sciences of mind*. Belknap.
- Tikhonov, A., & Yamshchikov, I. P. (2023). *Post Turing: Mapping the landscape of LLM Evaluation* (No. arXiv:2311.02049). arXiv. <https://doi.org/10.48550/arXiv.2311.02049>
- Tomasello, M. (2022). *The evolution of agency: Behavioral organization from lizards to humans*. The MIT Press.
- Torrance, S. (2014). Artificial consciousness and artificial ethics: Between realism and social relationism. *Philosophy & Technology*, 27(1), 9–29. <https://doi.org/10.1007/s13347-013-0136-5>
- Tunstall, L. (2022). *Understanding FLOPs-per-token estimates from OpenAI's scaling laws—Research [Research]*. Hugging Face Forums. <https://discuss.huggingface.co/t/understanding-flops-per-token-estimates-from-openais-scaling-laws/23133>
- Turing, A. (1950). Computing machinery and intelligence. *Mind*, LIX(236), 433–460. <https://doi.org/10.1093/mind/LIX.236.433>
- Ünsal, M. (2023). *DemoGPT: Autonomous AI Agent for Effortless App Creation* [Python]. <https://github.com/melih-unsal/DemoGPT> (Original work published 2023).
- Valdivia, A. (2024). The supply chain capitalism of AI: A call to (re)think algorithmic harms and resistance through environmental lens. *Information Communication & Society*, 0(0), 1–17. <https://doi.org/10.1080/1369118X.2024.2420021>
- Valmeekam, K., Sreedharan, S., Marquez, M., Hernandez, A. O., & Kambhampati, S. (2023). *On the Planning Abilities of Large Language Models—A Critical Investigation*. Thirty-seventh Conference on Neural Information Processing Systems. <https://openreview.net/forum?id=X6dEqXiSEW>
- Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J. B., Aali, A., Bluethgen, C., Pareek, A., Polacín, M., Reis, E. P., Seehofnerova, A., Rohatgi, N., Hosamani, P., Collins, W., Ahuja, N., Langlotz, C. P., Hom, J., Gatidis, S., Pauly, J., & Chaudhari, A. S. (2024). *Adapted Large Language Models Can Outperform Medical Experts in Clinical Text Summarization* (No. arXiv:2309.07430). arXiv. <http://arxiv.org/abs/2309.07430>.
- Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. MIT Press.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, 30. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a854aa-Abstract.html>
- Verbeek, P. P. (2008). Cyborg intentionality: Rethinking the phenomenology of human–technology relations. *Phenomenology and the Cognitive Sciences*, 7(3), 387–395.
- Waisberg, E., Ong, J., Masalkhi, M., Zaman, N., Sarker, P., Lee, A. G., & Tavakkoli, A. (2023). Meta smart glasses—Large Language models and the future for assistive glasses for individuals with vision impairments. *Eye (London, England)*, 1–3. <https://doi.org/10.1038/s41433-023-02842-z>
- Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. R. (2023). *A Survey on Large Language Model based Autonomous Agents*. arXiv. <https://arxiv.org/abs/2308.11432v2>.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (No. arXiv:2201.11903). arXiv. <https://doi.org/10.48550/arXiv.2201.11903>
- Weng, L. (2023). LLM Powered Autonomous Agents. *Lil'Log*. <https://lilianweng.github.io/posts/2023-06-23-agent/>
- Wolfram, S. (2023). What Is ChatGPT Doing... and Why Does It Work? [Personal]. *Stephen Wolfram Writings*. <https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/>
- Wu, S. (2024). Introducing devin, the first AI software engineer. *Cognition Labs*. <https://www.cognition-labs.com/introducing-devin>
- Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., & Wang, C. (2023). *AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework* [Jupyter Notebook]. <https://github.com/microsoft/autogen> (Original work published 2023).

- Wu, W., Morris, J. X., & Levine, L. (2024). *Do language models plan ahead for future tokens?* (No. arXiv:2404.00859). arXiv. <https://doi.org/10.48550/arXiv.2404.00859>
- Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., & Gui, T. (2023). *The Rise and Potential of Large Language Model Based Agents: A Survey* (No. arXiv:2309.07864). arXiv. <http://arxiv.org/abs/2309.07864>.
- Xiang, C. (2023). OpenAI Used Kenyan Workers Making \$2 an Hour to Filter Traumatic Content from ChatGPT. *Vice*. <https://www.vice.com/en/article/wxn3kw/openai-used-kenyan-workers-making-dollar-2-an-hour-to-filter-traumatic-content-from-chatgpt>
- Xiang, J., Tao, T., Gu, Y., Shu, T., Wang, Z., Yang, Z., & Hu, Z. (2023). *Language Models Meet World Models: Embodied Experiences Enhance Language Models* (No. arXiv:2305.10626). arXiv. <https://doi.org/10.48550/arXiv.2305.10626>
- Xie, T., Zhang, D., Chen, J., Li, X., Zhao, S., Cao, R., Hua, T. J., Cheng, Z., Shin, D., Lei, F., Liu, Y., Xu, Y., Zhou, S., Savarese, S., Xiong, C., Zhong, V., & Yu, T. (2024). *OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments* (No. arXiv:2404.07972). arXiv. <https://doi.org/10.48550/arXiv.2404.07972>
- Xie, J., Zhang, K., Chen, J., Yuan, S., Zhang, K., Zhang, Y., Li, L., & Xiao, Y. (2025). Revealing the Barriers of Language Agents in Planning. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 1872–1888). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.naacl-long.93>
- Yang, J., Jimenez, C. E., Wettig, A., Lieret, K., Yao, S., Narasimhan, K., & Press, O. (2024). *SWE-agent: Agent Computer Interfaces Enable Software Engineering Language Models*.
- Yao, S., Yu, D., Zhao, J., Shafraan, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). *Tree of Thoughts: Deliberate Problem Solving with Large Language Models* (No. arXiv:2305.10601). arXiv. <https://doi.org/10.48550/arXiv.2305.10601>
- Yildirim, I., & Paul, L. A. (2024). From task structures to world models: What do LLMs know? *Trends in Cognitive Sciences*, 28(5), 404–415. <https://doi.org/10.1016/j.tics.2024.02.008>
- Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). *HellaSwag: Can a Machine Really Finish Your Sentence?* (No. arXiv:1905.07830). arXiv. <https://doi.org/10.48550/arXiv.1905.07830>
- Zhou, A., Yan, K., Shlapentokh-Rothman, M., Wang, H., & Wang, Y. X. (2023a). *Language Agent Tree Search Unifies Reasoning Acting and Planning in Language Models* (No. arXiv:2310.04406). arXiv. <https://doi.org/10.48550/arXiv.2310.04406>
- Zhou, W., Jiang, Y. E., Li, L., Wu, J., Wang, T., Qiu, S., Zhang, J., Chen, J., Wu, R., Wang, S., Zhu, S., Chen, J., Zhang, W., Tang, X., Zhang, N., Chen, H., Cui, P., & Sachan, M. (2023b). *Agents: An Open-source Framework for Autonomous Language Agents* (No. arXiv:2309.07870). arXiv. <https://doi.org/10.48550/arXiv.2309.07870>
- Zhou, L., Schellaert, W., Martínez-Plumed, F., Moros-Daval, Y., Ferri, C., & Hernández-Orallo, J. (2024). Larger and more instructable Language models become less reliable. *Nature*, 634(8032), 61–68. <https://doi.org/10.1038/s41586-024-07930-y>
- Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2020). *Fine-Tuning Language Models from Human Preferences* (No. arXiv:1909.08593). arXiv. <https://doi.org/10.48550/arXiv.1909.08593>
- Ziemke, T. (2023). Understanding social robots: Attribution of intentional agency to artificial and biological bodies. *Artificial Life*, 29(3), 351–366. https://doi.org/10.1162/artl_a_00404
- Ziemke, T., & Lowe, R. (2009). On the role of emotion in embodied cognitive architectures: From organisms to robots. *Cognitive Computation*, 1(1), 104–117. <https://doi.org/10.1007/s12559-009-9012-0>

Authors and Affiliations

Xabier E. Barandiaran¹  · Lola S. Almendros² 

✉ Xabier E. Barandiaran
xabier.barandiaran@ehu.eus

Lola S. Almendros
lola.s.almendros@gmail.com

¹ IAS-Research Centre for Life, Mind and, Society Dept. Philosophy UPV/EHU, University of the Basque Country, Donostia, Spain

² Institute for Science and Technology Studies, University of Salamanca, Salamanca, Spain