On the Creativity of AI Agents

Giorgio Franceschelli
University of Bologna
giorgio.franceschelli@unibo.it

Mirco Musolesi
University College London and University of Bologna
m.musolesi@ucl.ac.uk

Abstract

Large language models (LLMs), particularly when integrated into agentic systems, have demonstrated human- and even superhuman-level performance across multiple domains. Whether these systems can truly be considered creative, however, remains a matter of debate, as conclusions heavily depend on the definitions, evaluation methods, and specific use cases employed. In this paper, we analyse creativity along two complementary macro-level perspectives. The first is a *functionalist* perspective, focusing on the observable characteristics of creative outputs. The second is an *ontological* perspective, emphasising the underlying processes, as well as the social and personal dimensions involved in creativity. We focus on LLM agents and we argue that they exhibit functionalist creativity, albeit not at its most sophisticated levels, while they continue to lack key aspects of ontological creativity. Finally, we discuss whether it is desirable for agentic systems to attain both forms of creativity, evaluating potential benefits and risks, and proposing pathways toward artificial creativity that can enhance human society.

1 Introduction

Recent years have seen the rise of agentic systems (e.g., [39, 52, 71]), i.e., autonomous or semi-autonomous software entities capable of reasoning, acting, and interacting with external environments[1]. These systems can plan multi-step solutions, invoke tools, retrieve information, and adapt their behaviour based on feedback, enabling them to tackle complex and dynamic tasks. At the core of many such systems lie large language models

[1]The study of (multi-agent) agent systems has a long tradition in Computer Science, which is often unfortunately overlooked (see, for example, [97, 110]). Early agent systems were primarily based on logic, and as a result, their capabilities were inherently limited. However, these systems laid important foundations for the advances we see today.

(LLMs), whose recent advancements are revolutionising several fields [17] due to their remarkable emergent capabilities [107]. LLMs provide the underlying reasoning and language understanding that make agentic behaviour possible. In particular, they can solve complex and even previously unseen tasks through prompt engineering [82] and in-context learning [19], i.e., by adjusting the input text or augmenting it with a few examples of expected outcomes. Building on these capabilities, modern agentic systems are typically structured around an often fine-tuned LLM[2] that serves as the basis for both *acting* (by generating textual commands to invoke tools) and *perceiving* (by incorporating retrieved information or environmental state into its input prompt).

These LLM agents have evolved beyond plain text generation into fully fledged acting systems capable of using tools and planning over complex tasks. For example, their capabilities have progressed from producing programming code in a single pass [8] to orchestrating more sophisticated workflows, in which tasks are decomposed into smaller sub-problems and addressed iteratively. Within these workflows, systems can debug and optimise generated code based on unit test performance, evaluate it at runtime, and return a solution only once all specified requirements are satisfied [48].

Their widespread use in tasks traditionally associated with creativity and innovation (not only in writing or coding, but also in academic research [98], scientific discovery [29], and problem-solving [103]) has sparked intense debate over whether they can truly be considered creative. Depending on the definition of creativity assumed [86, 105], the psychometric tests employed [43, 54], the use cases considered [66, 50], and the level of proficiency of human evaluators [79, 26], the research community is somewhat polarized between those arguing that LLMs and LLM agents possess (truly) creative capabilities and those who contest this view.

In this paper, we propose a dualistic framework that integrates both views, providing a coherent way to discuss creativity. The framework differentiates a *functionalist approach*, focusing on the observable traits of artefacts and ideas, from an *ontological approach*, which explores the essential nature of the creative process itself. Grounding our analysis in the mechanistic behaviour of LLM-based agentic systems, we explore how they can achieve specific forms of functionalist creativity, yet remain limited with respect to deeper, ontological aspects of creativity. We highlight the gaps in current agentic systems and consider whether autonomous agents might reach higher creative capacities, proposing research directions where enhancing creativity could simultaneously support societal benefits.

2 From Language Models to AI Agentic Systems

Large Language Models (LLMs) are autoregressive generative models built on the Transformer architecture [104]. They are described as large because they contain billions of

[2]We envisage the rise of agents based on multimodal foundation models. While this article focuses on LLM agents, the ideas presented here naturally extend to this broader class of systems.

parameters, and as language models because they are trained on vast text corpora (e.g., almost the entirety of Web content) to learn the structure and patterns of human language. Technically speaking, a $\boldsymbol{\theta}$-parametrised autoregressive language model is a probability distribution $p_{\boldsymbol{\theta}}(\mathbf{x})$ over a variable-length text sequence $\mathbf{x} = (x_1 \dots x_T)$, where T is the sequence length, and each token x_t is in a finite vocabulary $\mathcal{V}$ of size N. A token can be a word, a sub-word, or even a single character, a punctuation mark, or any other symbol or combination of symbols that appears in written form. The probability distribution is factorized as $p_{\boldsymbol{\theta}}(\mathbf{x}) = \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(x_t|\mathbf{x}_{<t})$, where $\mathbf{x}_{<t} = x_1 \dots x_{t-1}$. The language model is usually trained to maximise the likelihood of the true distribution $p^*(\mathbf{x})$ for any $\mathbf{x}$ from a reference dataset (the training set). In other words, given an input $\mathbf{x}_{<t}$, the model learns to approximate the probability of each token from $\mathcal{V}$ being x_t. By sampling from these learned next-token probability distributions, we can use the LLM to generate new sentences. Given a conditional input (the prompt) $\mathbf{z} = (z_1 \dots z_L)$, we can decode $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ as the continuation of $\mathbf{z}$, i.e., through the factorized representation $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) = \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(x_t|\mathbf{x}_{<\mathbf{t}}, \mathbf{z})$. Commonly, after maximising the likelihood of the training set, the language model is subject to an additional, shorter training (called fine-tuning) with reinforcement learning. Here, the $\boldsymbol{\theta}$-parametrised model is incentivized (or discouraged) to generate completions $\mathbf{x}$ that maximize (or minimize) human feedback [73] or other desired properties such as correctness, completeness, and appropriateness given the prompt $\mathbf{z}$ [35].

In essence, given a fragment of text, what an LLM does is to predict what is likely to come next according to its model of the statistics of human language [94], which is derived from training data and possibly tuned with human feedback. Then, the output returned to the user is sampled according to these predictions by means of one of several available sampling strategies, which might truncate certain tokens (e.g., [33, 47, 67]) as well as sharpen [76] or flatten [100] their likelihood but still leverage the predicted next-token probability distribution. However, LLMs are rarely used in isolation, particularly outside the research community. Instead, they are typically integrated into larger agentic systems, sometimes confusingly also referred to as "large language models", even though they are, strictly speaking, applications built around an LLM core [93].

LLM agents are *systems* where LLMs receive input in natural language from their environment and take autonomous actions to accomplish specific tasks [78]. Their basic building block is an LLM enhanced with augmentations such as retrieval (i.e., the LLM can generate its own search query, and the results are included in the next-step prompt) [59], tools (i.e., the system can call tools by writing programming code, including predefined APIs, that is executed in a container and whose response is returned as part of the next-step prompt) [81], and memory (i.e., the system stores and maintains an evolving memory state from prior interactions, and at each timestep the LLM input relies on additional, relevant information retrieved from such a memory in addition to the current observation and the original prompt) [49]. LLM agents can differ in terms of memory, reasoning, and acting tools [106]. For instance, reasoning can happen due to a few-shot

chain-of-thought [108] or a zero-shot "let's think step by step" chain-of-thought [55], and acting can even involve physical robots [5]. Finally, agentic LLMs can interact not only with the user, but also with each other, based on specific roles (as in simulated companies [80, 112]) or even in open-ended settings where the LLM agents do not have predetermined roles and interact semi-spontaneously, forming in-silico societies [74]. Regardless of the tools and techniques they can dispose of, these systems can be built as *workflows*, i.e., systems where LLMs and tools are orchestrated through predefined code paths and the task is decomposed a priori by the developer into fixed parallel or consecutive steps that involve LLMs in different ways; or as *agents*, i.e., systems where LLMs dynamically direct their own processes and tool usage based on environmental feedback to accomplish given tasks [90].

In other words, LLM-based agentic systems are programs where LLM outputs control the workflow. While this might sound amazing (and it partially is), in practice, this is done by running the very same LLM inference we mentioned before in a loop: at each timestep, the input of the LLM is made of the user prompt, the last observation from the environment, and a summary of the results of previous iterations; and at each timestep, the output of the LLM is usually composed of two parts: first, the LLM "reasoning" on what it should do (or, more correctly, predicting what is likely to come next after the prompt and some "thinking" keywords), and second, the LLM writing actions as executable code snippets that make use of one or more available tools (or, more correctly, predicting what is likely to come next after the prompt, the generated "reasoning", and some "acting" keywords). Then, the program merely executes it and appends its output to the LLM output as the "observation" from the environment. The final answer returned to the user is simply the last LLM output (which does not contain any code snippet, as it predicts it already has all necessary information) [85]. In this way, the LLM can solve tasks from a great variety of domains, provided that it has access to the appropriate tools, which definition plays a key role in the success of LLM-based agentic systems [111].

3 The Present of Creative AI Agents: A Theory-Grounded Taxonomy

The debate around creativity in AI is often fragmented and framed at different levels of abstraction. Opposite claims typically arise because there is no agreement upon what creativity means and how it should be evaluated in the context of AI and, particularly, agentic systems. Here, we aim to sort this out by separating creativity into two macro-levels: functionalist and ontological creativity.

3.1 A Functionalist Approach

According to Runco and Jaeger's standard definition [87], creativity requires *originality* and *effectiveness*. Ideas or products must be unusual, novel, or unique, but also useful or appropriate. Whatever name they take, these two dimensions, one requiring divergence from conventions and one requiring fitness to utility or value functions, are always present in any creativity definition and represent the two pillars around which studies in traditional computational creativity have been developed [84, 109]. A similar dichotomy is also present in Boden's definitions of creativity, which span from the ability to generate *novel* and *valuable* ideas [16] to the widely known tripartite definition of creativity as the ability to come up with ideas or artefacts that are *new*, *surprising*, and *valuable* [15].

While surprise is still a form of divergence, it brings to the discussion the role of timeliness and causality: something creative should be unexpected and non-obvious. As a confirmation of its relevance, according to most patent laws, an invention to qualify for protection must not only be useful in some practical sense (i.e., valuable) and not previously disclosed (i.e., novel), but also not obvious to someone ordinarily skilled in that field [9]. Similarly, a painting that is a mere variation of other artworks, and that remains within the realm of standard, reasonable paintings, is less surprising and thus less creative than one that introduces disruptive elements of originality. This suggests that creativity is not a binary property but rather something that can take different forms depending on the dimension of surprise. In particular, Boden identifies three forms of creativity. *Combinational* creativity involves making unfamiliar combinations of familiar ideas. *Exploratory* creativity involves the exploration of conceptual spaces to come up with new ideas within them that fit with them, where a conceptual space can be seen as the generative system underlying a domain that defines a certain range of possibilities [14]. *Transformational* creativity involves the transformation of such spaces by generating ideas that change the pre-existing style, so that previously inconceivable thoughts become possible [15].

The appealing nature of these creativity definitions, especially from the perspective of an AI researcher, is that they leverage observable properties. While, of course, there are plenty of examples in history of creative ideas not immediately recognised as such, effectiveness, originality, novelty, surprise, and value all concern external aspects of the creation: we just need the output (in the form of a concrete artefact or an abstract idea) to evaluate creativity under these dimensions, and similarly, we just need the output to identify what form of creativity we are dealing with. After all, observers can only evaluate the creativity of the final product and usually care only about how they feel about it. From a materialistic perspective, it is perfectly reasonable to find something creative if it appears to us as valuable (effective, appropriate, useful, or correct), novel (as something unconventional or, better, something we reasonably believe to be unconventional for the author), and to some degree surprising.

In the context of LLM-based agentic systems, we can have such a functionalist creativity at two different levels: in the final product, i.e., the outcome the agent returns to the

user, and in the single actions performed, i.e., the single steps the agent makes to arrive at the final product. It is possible to have creativity at one level while not having it at the other: the final product might be creative even if the single steps are somehow standard (this can especially happen with combinational creativity, where the single elements and ideas are common and it is only their final combination that is creative), and conversely, the final product might not be creative even if some of the single steps are creative (e.g., because their final impact is negligible; a chef may invent a new way to cut carrots, but this may not impact the stew they are going to serve). Of course, it is also likely that a creative product is caused by at least one creative action and that a creative action leads to a creative product (this is particularly true for transformational creativity, where single creative actions that break norms are required).

Discussing creativity in single actions might seem complicated, as their specific effects may not be easily detectable. However, it is not really different from discussing creativity in LLMs alone: they still choose how to act by predicting the most likely action name as the completion of the provided prompt, rather than deviating from norms. In other words, they strictly follow the generative system they possess for that domain. Unless fine-tuning has pushed them (directly or indirectly) to diverge from the learned human model, they will not be able to choose an action that breaks from previous understanding of the subject, even though they can navigate its boundaries and produce something combinational or exploratory.

Evaluating creativity in the final product seems easier, as it is apparent to the user and can thus be perceived as novel, surprising, and valuable. However, its degree of creativity is influenced by the individual actions and the "reasoning" behind them, as these are the aspects that shape future prompts and can therefore cause it to diverge more or less from the current space of solutions. The fact that they use predetermined tools, which of course adhere to our current conception of the domain, suggests that they cannot approach transformational creativity. However, recent agents are also capable of generating new tools [68]. In theory, this means that they can also create actions that significantly transform the conceptual space, leading to the purest forms of creativity. However, in practice, these tools are generated by an LLM, which is still a probabilistic model of human language, in a single action (whose creativity is discussed above), suggesting that it cannot generate something that deviates from what already exists in that way. This leads to a seeming conundrum: to overcome human limits, agents are provided with the ability to generate their own actions, but then they are constrained by being probabilistic models of human language and by sampling from within it, and thus remain subject to the same human limits (with the effect that they can only complement what the developer has programmed, filling a gap in the space rather than transforming it).

Indeed, LLM-based agentic systems are already capable of producing results that are valuable, novel, and surprising. However, these results are the product of interpolation (looking between the seen examples) or extrapolation (looking beyond the seen examples), but not yet of hyperpolation (transcending the seen examples) [72], which is instead

required to abstract from the defining dimensions of a domain and be able to alter or drop one or more of them, which is the essence of transformational creativity [16]. Similarly, these results are usually obtained via induction or deduction rather than abduction. Current LLMs cannot jump or break boundaries: they can execute the necessary steps to prove new theorems from established premises, but they are incapable of formulating original premises from sensed experience [113]. Again, creative abduction (the form of abduction where the law is invented *ex novo* to fit with the evidence in a more elegant and even aesthetically pleasant way [31]) is conducive to transformational creativity, as it is evident in revolutionary, path-breaking scientific discoveries [11, 56].

The mere fact that LLM agents cannot approach transformational creativity does not impact how creatively (in a combinational and exploratory sense) they can be, especially in scientific discovery. For instance, GPT-5.2 has recently derived a new formula in theoretical physics by spotting a pattern behind autonomously reduced expressions and positing a general formula for them [42]; a different version of the same model has also been capable of reaching the same formula and producing a formal proof by just "reasoning" through the problem [64]. Still, these results are closer to inductive than abductive reasoning, and to extrapolation than hyperpolation. How to effectively test for transformational creativity is an open, and potentially unsolvable, problem. Although tests such as superspace extrapolation exist, where training examples in an n-dimensional space are extended to an $n+1$-dimensional superspace encompassing them [63], it is hard to imagine how they could be systematically applied to the continuously expanding array of AI applications. Designing these tasks and specifying the corresponding $n+1$-dimensional superspaces already constitutes a highly creative, transformational challenge. Most likely, the purest form of creativity can only be assessed as an emergent property, e.g., because agents start showing the ability to invent a new programming paradigm, initiate an unprecedented artistic style, derive a novel non-Euclidean geometry, or achieve any other paradigm-changing result.

Finally, until now, we have partially neglected the role of the "reasoning" step, focusing more on the action selection and the final output. However, this step can play a significant role from a creativity perspective. For now, the most creative and least obvious outputs from LLMs and LLM agents are due to careful prompting: there is a strong correlation between output creativity and prompt creativity, and without a sufficiently specific and originality-facilitating prompt, the output tends to collapse to slops [45]. This is a known aspect of creativity and innovation, especially in science. As Einstein and Infeld put it, "the formulation of a problem is often more essential than its solution, which may be merely a matter of mathematical or experimental skill. To raise new questions, new possibilities, to regard old problems from a new angle requires creative imagination and marks real advances in science" [32]. Whether multiple, possibly divergent "reasoning" steps can create the conditions for a sufficiently specific and original prompt that leads to more creative outputs may represent the first and most ideal gate toward greater functionalist creativity.

3.2 An Ontological Approach

While the properties of the creative output and its effect on beholders and users are all that matter from a materialistic point of view, they represent only one side of the coin in creativity theories. Indeed, there is broad agreement among researchers that creativity should be studied and evaluated from perspectives beyond the mere product (e.g., [65, 101, 102]).

As summarised by Rhodes [83], three additional perspectives require attention: the process, the press, and the person. The creative *process* encompasses motivation, perception, learning, thinking, and communication [83], and requires both domain- and creativity-relevant skills [6]. Typically, once a problem is provided (by external or internal stimuli), the process involves a loop where one or multiple responses are generated and evaluated; if one response is found to be successful, the loop ends; otherwise, other responses are generated based on these findings (or the problem is revisited) [6]. On the other hand, the creative *press* refers to the relationship between a product and the influence its environment has upon it [83]. As we already discussed, products have to be accepted as creative by society; however, they should also be influenced by previously accepted works and, in turn, influence future ones. Thus, the environment should have a key role in how the product is shaped: it provides a culturally defined domain of action in which innovation is possible, and it contains a set of peers that evaluate whether the product is worthy of being promoted and preserved [25]. Finally, the *person* perspective acknowledges the personality, intellect, temperament, habits, attitude, value systems, and defence mechanisms of the producer [83], and requires the agent to exhibit intentionality [86] and purposes [38][3]. We have previously demonstrated that LLMs do not satisfy these requirements [36]. The question remains: do agentic systems perform any better?

Embedding the LLM as the core part of larger systems, as those discussed in Section 2, has arguably addressed two of the known limitations of LLMs under creativity theory. First, the agentic system can now evaluate its own output and decide whether to return it or not (e.g., [34, 53]). The model outputs the final solution only when it is convinced of it; otherwise, it can continue to gather new evidence from the environment, or to correct its approach via a different reasoning or the use of alternative tools. While, as already discussed, this planning and acting might not be creativity-oriented and not based on creativity-relevant skills, this *generate-then-evaluate* loop closely resembles the creative process described by Amabile [6]. Second, LLM-based agentic systems can perceive and interact with the environment in ways that go beyond the mere prompt engineering typical

[3] An interesting dimension of creativity, as proposed by Shanahan *et al.* in [95], lies in the concept of role-playing. In this approach, artificial agents simulate different personas, perspectives, or narrative roles, enabling them to explore creative possibilities from multiple angles. While this mechanism can enhance AI's capacity to generate novel ideas or solutions, it still fundamentally relies on human input, whether in defining the roles, setting the context, or guiding the objectives of the exercise. Going beyond this type of framing, in which the role-playing is driven by the AI system itself through intrinsic motivation, is an open area, which we are going to discuss in the next section.

of LLMs (even though the LLM still receives such information via automated prompts). In this way, they can acknowledge other agents, be influenced by their outputs, and deal with an evolving environment [69], addressing concerns regarding the lack of a social dimension in artificial creativity.

While these two properties are partially hard-coded and limited in scope, they nevertheless address the majority of the "easy problems" in AI creativity highlighted in [36], i.e., issues concerning exploration of different solutions and the production of outputs that diverge from conventions in constrained contexts. However, they still fail to resolve the "hard problems", which involve questions of deeper understanding, intentionality, and genuine transformation of the underlying conceptual space. In particular, it is possible to spot three main gaps, one for each creative perspective beyond product. First, they are not intrinsically motivated and work only on external inputs. However, motivation is an essential part of the creative process and usually comes from an intrinsic interest in the task (i.e., the task is interesting or enjoyable per se) [27], which also positively impacts the creativity of discovered solutions [7]. Similarly, solving discovered rather than externally presented problems is more conducive to creativity [40, 88]. Second, the influence of the environment and the impact of their own outcomes and other external products are typically limited to the specific episode and do not leave a trace on the LLM. While RAG-based systems can collect information about what has happened after the LLM deployment and solve its knowledge cut-off [37], and some agentic systems allow for the generation and preservation of new, artificial tools to expand the available toolset [68], such an updated information is still prompted to the same, old LLM, that will process it according to the same, old probability model of the human language. In summary, LLM-based agentic systems still lack the never-ending loop in which past experience shapes future experience; they still lack continual learning. Third, agentic systems do not possess personal traits and are not intentional agents. They lack both *liberty* (independence from controlling principles) and *agency* (capacity for intentional action) [51]; indeed, the system passively follows the programmed routine, and while the agent has the autonomy of deciding the next action, it cannot refuse to generate a response. At each timestep, the LLM output is the mere product of a probabilistic model upon an automatically formed input, and the final solution is the mere output at the last timestep, rather than the result of an intentional, personal, and experience-defined process.

All in all, there is a *fil rouge* linking these problems: the current lack of consciousness and self-awareness in artificial intelligence [20], despite evidence of minimal introspective capabilities [12, 22]. Indeed, being conscious is central for intentionality: to be intentional, a state or process must be thinkable or experienceable; and to be thinkable or experienceable, it must be, at least in principle, *consciously* thinkable or experienceable [92]. In addition, it has been recently argued that continual learning is necessary (but possibly not sufficient) for consciousness, and the lack of consciousness in LLMs may be due to the lack of continual learning [46]. Finally, intrinsically motivated actions should be volitional and experienced as congruent and self-endorsed [28]. In other words, intelligence

and sentience are two very distinct properties [13, 58]. While the first may be sufficient to *do* something creative, both are necessary to *be* creative, at least under more rigorous, stringent philosophical definitions.

4 The Future of AI Creative Agents: To Infinity and Beyond

While current agentic systems may fail to approach creativity under certain perspectives and may be partially limited under others, their creative achievements in recent years have been remarkable. For example, LLMs have moved from struggling with basic maths [70] to passing university-level exams [18], and now, having been embedded in larger agentic systems, they can even participate in scientific research [91]. At this pace, it seems plausible to predict a future in which agents can reach the highest forms of functionalist creativity and may even be considered capable of ontological creativity. However, before proceeding down these paths, we should ask ourselves: where should we aim? Is a never-ending progression toward fully human-like creativity desirable after all?

4.1 Transformational AI Creativity as Augmentation, Not Substitution

Historically, reaching transformational (functionalist) creativity does not doom current solutions to disappearance; rather, the new $n + 1$-dimensional space typically allows novel and past styles of thinking to coexist. The birth of abstract painting did not cause figurative art to disappear; non-Euclidean geometries did not prevent Euclidean geometry from remaining the most studied; new programming paradigms and languages did not completely replace previous ones. On the contrary, transformationally creative artefacts can be inspirational sources for other creative outputs (as the modified space can now be explored in different ways, thus being conducive to more creativity [15]), and they might even unlock new understandings of life and science that were previously subject to traditional assumptions, as happened several times in history, from Copernican heliocentrism to Einstein's general theory of relativity. In this sense, agentic systems might, for example, overcome constraints dictated by human (physical or mental) limitations.

Moreover, functionalist creativity can also lead to outcomes that are orthogonal or complementary to our creativity. For instance, Nature is, in a sense, functionally creative: it produces artefacts (such as shells, leaves, plants, even landscapes) that can be perceived as valuable, novel, and surprising, without being *truly* creative from an agentic perspective (as its underlying process is non-agentic [75]). Artificial creativity might reach similar levels of utility in terms of aesthetics, profitability, and even research: just like the study of Nature's artefacts reveals something about the world we live in, the study of creative AI artefacts can reveal something about AI and humans as well.

In addition, moving generative AI and agentic systems from combinational (or no creativity at all) to deeper forms of creativity, where the outputs are less derivative and

more divergent, can reasonably position AI outputs beside rather than in place of human outputs. Indeed, creators and researchers might use creative agents to expand and complement their work, while non-creative, replicative systems may be seen as shortcuts (in terms of time and cost) to human results. Returning less derivative products and generating more uncommon outputs may therefore reduce ethical and legal risks, preserving human intellectual property [23] as well as the established human roles in creative and innovative domains [1].

Because of these considerations, we argue that pursuing functionalist creativity is a promising direction for next-generation LLM-based agentic systems. Encouraging greater, yet valuable, divergence at both the learning and inference levels, alongside enhancing reasoning and tool use with creative capabilities, can represent an important first step forward. Since hyperpolation may be reached via abduction or, conversely, may serve as a mathematical model for explaining abduction [72], addressing one may also help resolve the other. Surpassing benchmarks and test suites that favour repetitive and derivative outputs, and developing new evaluative approaches, is another timely research direction.

4.2 The Challenging Nature of Intrinsic Motivation in AI Creativity

Reaching ontological creativity would require a different set of skills and properties to be possessed by LLM-based agentic systems. In particular, in Section 3.2 we identify three main gaps of current agentic systems under the ontological approach: lack of intrinsic motivation; lack of experience-based continual learning; and lack of intentionality and personality.

Intrinsic motivation is not a novel topic in AI research, especially around reinforcement learning [99]; several methods already exist to incorporate a notion of intrinsic motivation and curiosity during learning, to push the agent to learn deeper and further [57]. However, this is an approximate, synthetic version of intrinsic motivation, in which agents are merely incentivised to seek solutions that not only maximise environmental rewards but also possess desirable characteristics (e.g., those that favour learning or cover under-explored regions of the solution space [41]). The sort of (intrinsic) motivation considered by creativity theories is instead about having an intrinsic interest in the task at hand, or even choosing and reframing the task to find it more enjoyable [24]. The approximated version already studied in RL can undoubtedly have a positive impact on agentic LLMs, helping them find multiple ways of solving a problem or pushing them to explore multiple strategies during reasoning.

Having systems that can reframe problems into something more interesting, in the sense of being challenging but still feasible, can enhance research and innovation in several ways. On the other hand, the most authentic form of intrinsic motivation may have disruptive consequences, as an enjoyable task can also be useless or too divergent from what we are asking for, potentially allowing agents to escape human control.

4.3 Continual Learning and Self-Improvement in Creative AI Agents

Continual learning has been receiving increasing attention from AI researchers [44]. A central challenge lies in incorporating novel information while correcting or updating prior knowledge, a topic that has become a major focus of current research. Most existing approaches concentrate on extending training to new, online data while preserving previously acquired skills [96]. However, advancing beyond the knowledge contained in the original training set remains a significant and rapidly growing area of investigation[4].

As of now, the learning scheme remains fixed, and the same goes for the parameters that govern it. As a recent alternative, self-improving agents leverage the self-referential nature of these systems, which can be provided with tools to analyse, modify, and evaluate themselves, and can continually improve and simultaneously learn how to improve [114]; however, they are limited to computable tasks that allow for empirical validation (which is arguably hard in creative domains), and base their improvements exclusively upon such evaluation. On the other hand, human-like continual learning is always influenced by personal experience, and each new piece of information is acquired differently depending on its emotional impact, individual interests, and internal goals. The narrower version of continual learning can positively affect human creativity, as it can help LLM agents develop more effective interaction styles, align with user preferences, and build coherent artistic and research trajectories, in which each new creation evolves from previous ones. However, an agent that actively learns through a range of different schemes, selecting the most appropriate for each new experience, and that can set its own goals and develop its own reward functions, may diverge too far and cease to be a useful tool for creativity.

While humans rely on intrinsically motivated, experience-driven learning shaped by emotion, embodiment, and long-term goals, existing LLM agents operate through optimisation processes, reinforcement signals, and externally defined objectives. Currently, their behaviour reflects increasingly sophisticated forms of goal-directed optimisation and adaptation. However, the current research on more autonomous, tool-using, and self-improving systems suggests that, as models gain the ability to select learning strategies, update their behaviour over time, and pursue complex objectives, they may exhibit forms of functional autonomy that reduce direct human control. Empirical studies already show that LLM agents can match or exceed human performance in specific knowledge-intensive tasks, including coding, design ideation, and scientific problem-solving, particularly when augmented with retrieval, planning, and feedback mechanisms [21].

[4]Buzz Lightyear's iconic catchphrase "To infinity and beyond" [77] was selected as the title for this section, because it offers a compelling metaphor for the potential of AI creativity based on intrinsic motivation and continual learning. Just as the character aspires to transcend known limits of space, AI systems can push the boundaries of imagination, generating ideas, concepts, and artefacts that often surpass conventional human expectations. This represents one of the most interesting current research directions in this space.

4.4 Multi-Agent AI Systems and Creativity

Multi-agent AI systems introduce a further dimension to computational creativity by enabling interaction, collaboration, and competition among multiple autonomous agents [89]. Rather than treating creativity as an isolated property of a single model, we can consider it as an emergent phenomenon arising from distributed processes, where diverse agents contribute distinct perspectives, strategies, and generative biases. In creative settings, interacting agents can facilitate idea generation through constructive brainstorming and debate. Agents may propose, critique, refine, or combine outputs, leading to iterative improvement beyond what a single system could typically achieve [61]. This is particularly relevant in tasks requiring exploration of large or structured solution spaces, where diversity of thought and possibly contrasting hypothesis generation are beneficial. In fact, in such contexts, creativity may emerge not only from individual generation but also from negotiation, coordination across and even competition among agents.

Moreover, competitive multi-agent setups can stimulate creative divergence. When agents are incentivised to outperform one another under shared or partially conflicting objectives, they may explore more unconventional or high-risk solutions [115]. Cooperative settings, in contrast, tend to promote convergence and refinement, potentially improving coherence and usability of outputs [60]. The balance between cooperation and competition therefore plays a key role in shaping the nature of generated artefacts [30]. Indeed, future progress in AI creativity may depend not only on improving individual models, but also on designing effective ecosystems of interacting AI agents.

4.5 AI Agentic Systems and the Transformation of Creative Work

Given their scalability and computational efficiency, such systems could significantly impact creative and knowledge-based professions. Economic analyses of automation and generative AI adoption indicate potential displacement or transformation of roles in art, research, and invention, alongside productivity gains [3, 2, 4]. At the same time, psychological research highlights the importance of creativity for identity, meaning, and well-being, suggesting that reduced human centrality in creative processes could have non-trivial psychological effects. In other words, this shift could carry enormous economic (i.e., occupational) implications, alongside significant psychological effects stemming from a diminished sense of control over what has long been considered a defining human trait: the ability to create [10].

5 Conclusion

LLMs, and especially LLM agents, are increasingly employed by researchers and practitioners to tackle a wide range of tasks traditionally associated with creativity. Yet, the question of whether these systems are genuinely creative remains contentious, hinging on

how creativity is defined and the perspective from which it is assessed. In this paper, we have proposed a dualistic framework designed to reconcile these differing viewpoints and provide a principled foundation for evaluating creativity in both LLMs and LLM-based agentic systems. In particular, we have conceptualised creativity at two macro levels: functionalist creativity, which pertains to the observable properties of generated artefacts or ideas, and ontological creativity, which concerns the underlying characteristics of the generative process as well as the personal and social conditions necessary for creative acts. We have then applied these definitions to examine current agentic LLMs, arguing that they exhibit functionalist creativity to some extent but still lack the hyperpolation and abductive reasoning necessary for genuine transformational creativity. Furthermore, while embedding LLMs within larger, agentic systems endows them with certain capacities relevant to ontological creativity, they remain deficient in three fundamental areas: intrinsic motivation, continual learning, and intentionality and authenticity.

Although current capabilities remain limited, recent advances indicate that many of these constraints could be overcome in the coming years. Some of these forthcoming developments (particularly those enhancing functionalist creativity) may serve as powerful tools for human creators. Conversely, other advances could pose significant risks to creative practice. We argue that fostering discussions around creativity is essential for shaping the future of AI in creative domains across legal [23], ethical [116], and scientific [62] dimensions, providing a shared framework to determine both the directions we should pursue and the strategies needed to safeguard the central role of *homo faber* [10].

References

[1] S. AbuMusab. Generative AI and human labor: who is replaceable? *AI & SOCIETY*, 39(6):3051–3053, 2024.

[2] D. Acemoglu. The simple macroeconomics of AI. *Economic Policy*, 40(121):13–58, 2025.

[3] D. Acemoglu and P. Restrepo. Artificial intelligence, automation, and work. In *The Economics of Artificial Intelligence: An Agenda*, pages 197–236. University of Chicago Press, 2018.

[4] D. Acemoglu, D. Kong, and A. Ozdaglar. AI, human cognition and knowledge collapse. Technical report, National Bureau of Economic Research, 2026.

[5] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao,

K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan, and A. Zeng. Do as i can, not as i say: Grounding language in robotic affordances. In *Proc. of the 6th Conference on Robot Learning (CoRL'22)*, 2022.

[6] T. M. Amabile. The social psychology of creativity: A componential conceptualization. *Journal of Personality and Social Psychology*, 45(2):357–376, 1983.

[7] T. M. Amabile. Motivation and creativity: Effects of motivational orientation on creative writers. *Journal of Personality and Social Psychology*, 48(2):393–399, 1985.

[8] J. Austin, A. Odena, M. Nye, M. Bosma, H. Michalewski, D. Dohan, E. Jiang, C. Cai, M. Terry, Q. Le, and C. Sutton. Program synthesis with large language models, 2021. arXiv:2108.07732 [cs.PL].

[9] J. H. Barton. Non-obviousness. *IDEA: The Journal of Law and Technology*, 43:475–508, 2003.

[10] H. Bergson. *Creative evolution*. Henry Holt and Company, 1911.

[11] T. M. Bertilsson. The elementary forms of pragmatism: On different types of abduction. *European Journal of Social Theory*, 7(3):371–389, 2004.

[12] F. J. Binder, J. Chua, T. Korbak, H. Sleight, J. Hughes, R. Long, E. Perez, M. Turpin, and O. Evans. Looking inward: Language models can learn about themselves by introspection. In *Proc. of the 13th International Conference on Learning Representations (ICLR'25)*, 2025.

[13] J. Birch. *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*. Oxford University Press, 2024.

[14] M. A. Boden. What is creativity? In *Dimensions of Creativity*, pages 75–117. The MIT Press, 1994.

[15] M. A. Boden. *The Creative Mind: Myths and Mechanisms*. Routledge, 2003.

[16] M. A. Boden. Computer Models of Creativity. *AI Magazine*, 30(3):23–34, 2009.

[17] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Quincy Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W.

Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang. On the opportunities and risks of foundation models, 2021. arXiv:2108.07258 [cs.LG].

[18] B. Borges, N. Foroutan, D. Bayazit, A. Sotnikova, S. Montariol, T. Nazaretsky, M. Banaei, A. Sakhaeirad, P. Servant, S. P. Neshaei, J. Frej, A. Romanou, G. Weiss, S. Mamooler, Z. Chen, S. Fan, S. Gao, M. Ismayilzada, D. Paul, P. Schwaller, S. Friedli, P. Jermann, T. Käser, A. Bosselut, E. G. Consortium, and E. D. Consortium. Could ChatGPT get an engineering degree? Evaluating higher education vulnerability to AI assistants. *Proceedings of the National Academy of Sciences*, 121(49): e2414955121, 2024.

[19] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NIPS'20)*, 2020.

[20] P. Butlin, R. Long, E. Elmoznino, Y. Bengio, J. Birch, A. Constant, G. Deane, S. M. Fleming, C. Frith, X. Ji, R. Kanai, C. Klein, G. Lindsay, M. Michel, L. Mudrik, M. A. K. Peters, E. Schwitzgebel, J. Simon, and R. VanRullen. Consciousness in artificial intelligence: Insights from the science of consciousness, 2023. arXiv:2308.08708 [cs.AI].

[21] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. Ponde de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. Petroski Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. Hebgen Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba. Evaluating large language models trained on code, 2021. arXiv:2107.03374 [cs.LG].

[22] I. M. Comsa and M. Shanahan. Does it make sense to speak of introspection in large language models?, 2025. arXiv:2506.05068 [cs.CL].

[23] A. F. Cooper and J. Grimmelmann. The files are in the computer: Copyright, memorization, and generative AI. *Chicago-Kent Law Review*, 100:141, 2025.

[24] R. S. Crutchfield. Conformity and creative thinking. In *Contemporary approaches to creative thinking: A symposium held at the University of Colorado*, pages 120–140. Atherton Press, 1962.

[25] M. Csikszentmihalyi. Society, culture, and person: A systems view of creativity. In *The Nature of Creativity: Contemporary Psychological Perspectives*, pages 325–339. Cambridge University Press, 1988.

[26] E. Davis. ChatGPT's poetry is incompetent and banal: A discussion of (Porter and Machery, 2024), 2024. `https://cs.nyu.edu/~davise/papers/GPT-Poetry.pdf` [Accessed: 2026-04-08].

[27] E. L. Deci and R. M. Ryan. *Intrinsic Motivation and Self-Determination in Human Behavior*. Springer, 1985.

[28] E. L. Deci, R. M. Ryan, P. P. Schultz, and C. P. Niemiec. Being aware and functioning fully. In *Handbook of Mindfulness: Theory, Research, and Practice*. The Guilford Press, 2015.

[29] Y. Du, B. Yu, T. Liu, T. Shen, J. Chen, J. G. Rittig, K. Sun, Y. Zhang, Z. Song, B. Zhou, C. Masschelein, Y. Wang, H. Wang, H. Jia, C. Zhang, H. Zhao, M. Ester, T. Head-Gordon, C. P. Gomes, H. Sun, C. Duan, P. Schwaller, and W. Jin. Accelerating scientific discovery with autonomous goal-evolving agents, 2025. arXiv:2512.21782 [cs.AI].

[30] E. A. Duéñez-Guzmán, S. Sadedin, J. X. Wang, K. R. McKee, and J. Z. Leibo. A social path to human-like artificial intelligence. *Nature Machine Intelligence*, 5(11): 1181–1188, 2023.

[31] U. Eco. Horns, hooves, insteps: Some hypotheses on three types of abduction. In *The Sign of Three: Dupin, Holmes, Peirce*, pages 198–220. Indiana University Press, 1983.

[32] A. Einstein and L. Infeld. *The Evolution of Physics: The Growth of Ideas from Early Concepts to Relativity and Quanta*. Cambridge University Press, 1938.

[33] A. Fan, M. Lewis, and Y. Dauphin. Hierarchical neural story generation. In *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (ACL'18)*, 2018.

[34] T. Feng, T. H. Trinh, G. Bingham, D. Hwang, Y. Chervonyi, J. Jung, J. Lee, C. Pagano, S. hyun Kim, F. Pasqualotto, S. Gukov, J. N. Lee, J. Kim, K. Hou, G. Ghiasi, Y. Tay, Y. Li, C. Kuang, Y. Liu, H. Lin, E. Z. Liu, N. Nayakanti, X. Yang, H.-T. Cheng, D. Hassabis, K. Kavukcuoglu, Q. V. Le, and T. Luong. Towards autonomous mathematics research, 2026. arXiv:2602.10177 [cs.LG].

[35] G. Franceschelli and M. Musolesi. Reinforcement learning for generative AI: State of the art, opportunities and open research challenges. *Journal of Artificial Intelligence Research*, 79:417–446, 2024.

[36] G. Franceschelli and M. Musolesi. On the creativity of large language models. *AI & Society*, 40:3785–3795, 2025.

[37] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang. Retrieval-augmented generation for large language models: A survey, 2024. arXiv:2312.10997 [cs.CL].

[38] B. Gaut. The philosophy of creativity. *Philosophy Compass*, 5(12):1034–1046, 2010.

[39] Gemini Team and Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. arXiv:2507.06261 [cs.CL].

[40] J. W. Getzels. Problem-finding and the inventiveness of solutions. *The Journal of Creative Behavior*, 9(1):12–18, 1975.

[41] J. Gottlieb, P.-Y. Oudeyer, M. Lopes, and A. Baranes. Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences*, 17(11):585–593, 2013.

[42] A. Guevara, A. Lupsasca, D. Skinner, A. Strominger, and K. Weil. Single-minus gluon tree amplitudes are nonzero, 2026. arXiv:2602.12176 [hep-th].

[43] E. E. Guzik, C. Byrge, and C. Gilde. The originality of machines: AI takes the Torrance Test. *Journal of Creativity*, 33(3):100065, 2023.

[44] R. Hadsell, D. Rao, A. A. Rusu, and R. Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences*, 24(12):1028–1040, 2020.

[45] E. Hoel. Bits in, bits out, 2026. `https://www.theintrinsicperspective.com/p/bits-in-bits-out` [Accessed: 2026-04-08].

[46] E. Hoel. A disproof of large language model consciousness: The necessity of continual learning for consciousness, 2026. arXiv:2512.12802 [q-bio.NC].

[47] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration. In *Proc. of the 8th International Conference on Learning Representations (ICLR'20)*, 2020.

[48] Y. Hu, Q. Zhou, Q. Chen, X. Li, L. Liu, D. Zhang, A. Kachroo, T. Oz, and O. Tripp. QualityFlow: An agentic workflow for program synthesis controlled by LLM quality checks, 2025. arXiv:2501.17167 [cs.SE].

[49] Y. Hu, S. Liu, Y. Yue, G. Zhang, B. Liu, F. Zhu, J. Lin, H. Guo, S. Dou, Z. Xi, S. Jin, J. Tan, Y. Yin, J. Liu, Z. Zhang, Z. Sun, Y. Zhu, H. Sun, B. Peng, Z. Cheng, X. Fan, J. Guo, X. Yu, Z. Zhou, Z. Hu, J. Huo, J. Wang, Y. Niu, Y. Wang, Z. Yin, X. Hu, Y. Liao, Q. Li, K. Wang, W. Zhou, Y. Liu, D. Cheng, Q. Zhang, T. Gui, S. Pan, Y. Zhang, P. Torr, Z. Dou, J.-R. Wen, X. Huang, Y.-G. Jiang, and S. Yan. Memory in the age of AI agents, 2026. arXiv:2512.13564 [cs.CL].

[50] M. Ismayilzada, C. Stevenson, and L. van der Plas. Evaluating creative short story generation in humans and large language models. In *Proc. of the 16th International Conference on Computational Creativity (ICCC'25)*, 2025.

[51] A. Issak. Artistic autonomy in AI art. In *Proc. of NIPS'21 Machine Learning for Creativity and Design Workshop*, 2021.

[52] Kimi Team. Kimi K2: Open agentic intelligence, 2026. arXiv:2507.20534 [cs.LG].

[53] D. Knuth. Claude's cycles, 2026. `https://cs.stanford.edu/~knuth/papers/claude-cycles.pdf` [Accessed: 2026-04-08].

[54] M. Koivisto and S. Grassini. Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific Reports*, 13(1):13601, 2023.

[55] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. In *Proc. of the 36th International Conference on Neural Information Processing Systems (NIPS'22)*, 2022.

[56] T. S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.

[57] P. Ladosz, L. Weng, M. Kim, and H. Oh. Exploration in deep reinforcement learning: A survey. *Information Fusion*, 85:1–22, 2022.

[58] S. Lavelle. The machine with a human face: From artificial intelligence to artificial sentience. In *Advanced Information Systems Engineering Workshops (CAiSE'20)*, 2020.

[59] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NIPS'20)*, 2020.

[60] G. Li, H. A. Al Kader Hammoud, H. Itani, D. Khizbullin, and B. Ghanem. CAMEL: communicative agents for "mind" exploration of large language model society. In *Proc. of the 37th International Conference on Neural Information Processing Systems (NIPS'23)*, 2023.

[61] Y.-C. Lin, K.-C. Chen, Z.-Y. Li, T.-H. Wu, T.-H. Wu, K.-Y. Chen, H.-y. Lee, and Y.-N. Chen. Creativity in LLM-based multi-agent systems: A survey. In *Proc. of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP'25)*, 2025.

[62] L.-C. Lu, M. Liu, P. C. Lu, Y. Tian, S.-H. Sun, and N. Peng. Rethinking creativity evaluation: A critical analysis of existing creativity evaluations. In *Proc. of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL'26)*, 2026.

[63] C. G. Lucas, D. Sterling, and C. Kemp. Superspace extrapolation reveals inductive biases in function learning. In *Proc. of the 34th Annual Meeting of the Cognitive Science Society (CogSci'12)*, 2012.

[64] A. Lupsasca. GPT-5.2 derives a new result in theoretical physics, 2026. `https://openai.com/index/new-result-theoretical-physics/` [Accessed: 2026-04-08].

[65] D. W. MacKinnon. Creativity: A multi-faceted phenomenon. In *Creativity: A discussion at the Nobel conference*, page 17–32. North-Holland Publishing Company, 1970.

[66] G. Marco, L. Rello, and J. Gonzalo. Small language models can outperform humans in short creative writing: A study comparing SLMs with humans and LLMs. In *Proc. of the 31st International Conference on Computational Linguistics (COLING'25)*, 2025.

[67] N. N. Minh, A. Baker, C. Neo, A. G. Roush, A. Kirsch, and R. Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent LLM outputs. In *Proc. of the 13th International Conference on Learning Representations (ICLR'25)*, 2025.

[68] D. Nguyen, V. D. Lai, S. Yoon, R. A. Rossi, H. Zhao, R. Zhang, P. Mathur, N. Lipka, Y. Wang, T. Bui, F. Dernoncourt, and T. Zhou. Dynasaur: Large language agents beyond predefined actions. In *Proc. of the 2nd Conference on Language Modeling (COLM'25)*, 2025.

[69] E. Nisioti, S. Risi, I. Momennejad, P.-Y. Oudeyer, and C. Moulin-Frier. Collective innovation in groups of large language models. In *Proc. of the 2024 Conference on Artificial Life (ALIFE'24)*, 2024.

[70] R. Nogueira, Z. Jiang, and J. Lin. Investigating the limitations of transformers with simple arithmetic tasks, 2021. arXiv:2102.13019 [cs.CL].

[71] OpenAI. OpenAI o3 and o4-mini system card, 2025. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf [Accessed: 2026-04-08].

[72] T. Ord. Interpolation, extrapolation, hyperpolation: Generalising into new dimensions, 2024. arXiv:2409.05513 [cs.LG].

[73] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NIPS'22)*, 2022.

[74] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proc. of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST'23)*, 2023.

[75] E. S. Paul and D. Stokes. Attributing creativity. In *Creativity and Philosophy*, pages 193–209. Routledge, 2018.

[76] M. Peeperkorn, T. Kouwenhoven, D. Brown, and A. Jordanous. Is temperature the creativity parameter of large language models? In *Proc. of the 15th International Conference on Computational Creativity (ICCC'24)*, 2024.

[77] Pixar Animation Studios. Toy Story, 1995.

[78] A. Plaat, M. van Duijn, N. van Stein, M. Preuss, P. van der Putten, and K. Joost Batenburg. Agentic large language models, a survey. *Journal of Artificial Intelligence Research*, 84:29:1–74, 2025.

[79] B. Porter and E. Machery. AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports*, 14(1):26133, 2024.

[80] C. Qian, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li, C. Yang, W. Chen, Y. Su, X. Cong, J. Xu, D. Li, Z. Liu, and M. Sun. ChatDev: Communicative agents for software development. In *Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL'24)*, 2024.

[81] Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu, Y. Lin, X. Cong, X. Tang, B. Qian, S. Zhao, L. Hong, R. Tian, R. Xie, J. Zhou, M. Gerstein, dahai li, Z. Liu, and M. Sun. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *Proc. of the 12th International Conference on Learning Representations (ICLR'24)*, 2024.

[82] L. Reynolds and K. McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI'21)*, 2021.

[83] M. Rhodes. An analysis of creativity. *The Phi Delta Kappan*, 42(7):305–310, 1961.

[84] G. Ritchie. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines*, 17:76–99, 2007.

[85] A. Roucher, Merve, and T. Wolf. Introducing smolagents, a simple library to build agents, 2024. `https://huggingface.co/blog/smolagents` [Accessed: 2026-04-08].

[86] M. A. Runco. Updating the standard definition of creativity to account for the artificial creativity of AI. *Creativity Research Journal*, 37(1):1–5, 2025.

[87] M. A. Runco and G. J. Jaeger. The standard definition of creativity. *Creativity Research Journal*, 24(1):92–96, 2012.

[88] M. A. Runco and S. M. Okuda. Problem discovery, divergent thinking, and the creative process. *Journal of Youth and Adolescence*, 17(3):211–220, 1988.

[89] R. Saunders and O. Bown. Computational social creativity. *Artificial Life*, 21(3): 366–378, 2015.

[90] E. Schluntz and B. Zhang. Building effective agents, 2024. `https://www.anthropic.com/engineering/building-effective-agents` [Accessed: 2026-04-08].

[91] S. Schmidgall, Y. Su, Z. Wang, X. Sun, J. Wu, X. Yu, J. Liu, M. Moor, Z. Liu, and E. Barsoum. Agent laboratory: Using LLM agents as research assistants. In *Findings of the Association for Computational Linguistics (EMNLP'25)*, 2025.

[92] J. R. Searle. Consciousness, unconsciousness and intentionality. *Philosophical Issues*, 1:45–66, 1991.

[93] M. Shanahan. Still "talking about large language models": Some clarifications, 2024. arXiv:2412.10291 [cs.CL].

[94] M. Shanahan. Talking about large language models. *Commun. ACM*, 67(2):68–79, 2024.

[95] M. Shanahan, K. McDonell, and L. Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023.

[96] H. Shi, Z. Xu, H. Wang, W. Qin, W. Wang, Y. Wang, Z. Wang, S. Ebrahimi, and H. Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 58(5), 2025.

[97] Y. Shoham and K. Leyton-Brown. *Multiagent Systems: Algorithmic, Game-theoretic, and Logical Foundations*. Cambridge University Press, 2008.

[98] C. Si, D. Yang, and T. Hashimoto. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In *Proc. of the 13th International Conference on Learning Representations (ICLR'25)*, 2025.

[99] S. Singh, A. G. Barto, and N. Chentanez. Intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS'04)*, 2004.

[100] Y. Song, G. Wang, S. Li, and B. Y. Lin. The good, the bad, and the greedy: Evaluation of LLMs should not ignore non-determinism. In *Proc. of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL'25)*, 2025.

[101] R. J. Sternberg and T. I. Lubart. An investment theory of creativity and its development. *Human Development*, 34(1):1–31, 1991.

[102] T. Z. Tardif and R. J. Sternberg. What do we know about creativity? In *The nature of creativity*, page 429–440. Cambridge University Press, 1988.

[103] Y. Tian, A. Ravichander, L. Qin, R. Le Bras, R. Marjieh, N. Peng, Y. Choi, T. Griffiths, and F. Brahman. MacGyver: Are large language models creative problem solvers? In *Proc. of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'24)*, 2024.

[104] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS'17)*, 2017.

[105] H. Wang, J. Zou, M. Mozer, A. Goyal, A. Lamb, L. Zhang, W. J. Su, Z. Deng, M. Q. Xie, H. Brown, and K. Kawaguchi. Can AI be as creative as humans?, 2024. arXiv:2401.01623 [cs.AI].

[106] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J. Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):18635, 2024.

[107] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

[108] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proc. of the 36th International Conference on Neural Information Processing Systems (NIPS'22)*, 2022.

[109] G. A. Wiggins. Searching for computational creativity. *New Generation Computing*, 24:209–222, 2006.

[110] M. Wooldridge. *An Introduction to Multiagent Systems*. John Wiley & Sons, 2009.

[111] B. Wu, A. Jones, A. Renault, H. Tay, J. Noble, N. McCandlish, N. Picard, S. Jiang, and Claude Developer Platform team. Introducing advanced tool use on the claude developer platform, 2025. `https://www.anthropic.com/engineering/advanced-tool-use` [Accessed: 2026-04-08].

[112] M. Wu, J. Xu, Y. Yuan, G. Haffari, L. Wan, W. Luo, and K. Zhang. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *Transactions of the Association for Computational Linguistics*, 13:901–922, 2025.

[113] T. Zahavy. LLMs can't jump, 2026. `https://philsci-archive.pitt.edu/28024/`.

[114] J. Zhang, B. Zhao, W. Yang, J. Foerster, J. Clune, M. Jiang, S. Devlin, and T. Shavrina. Hyperagents, 2026. arXiv:2212.14392 [cs.LG].

[115] R. Zhang and S. Eger. LLM-based multi-agent poetry generation in non-cooperative environments. *Journal of Language Modelling*, 13(2):261–318, 2026.

[116] S. Zhu, Z. Wang, Y. Zhuang, Y. Jiang, M. Guo, X. Zhang, and Z. Gao. Exploring the impact of ChatGPT on art creation and collaboration: Benefits, challenges and ethical implications. *Telematics and Informatics Reports*, 14:100138, 2024.