Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PERSPECTIVE

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2024, 3, 1933

Towards a science exocortex

Kevin G. Yager *

Artificial intelligence (AI) methods are poised to revolutionize intellectual work, with generative AI enabling automation of text analysis, text generation, and simple decision making or reasoning. The impact to science is only just beginning, but the opportunity is significant since scientific research relies fundamentally on extended chains of cognitive work. Here, we review the state of the art in agentic AI systems, and discuss how these methods could be extended to have even greater impact on science. We propose the development of an exocortex, a synthetic extension of a person's cognition. A science exocortex could be designed as a swarm of AI agents, with each agent individually streamlining specific researcher tasks, and whose inter-communication leads to emergent behavior that greatly extend the researcher's cognition and volition.

Received 24th June 2024
Accepted 15th August 2024

DOI: 10.1039/d4dd00178h

rsc.li/digitaldiscovery

1 Introduction

Artificial intelligence and machine-learning (AI/ML) methods are having growing impact across a wide range of fields, including the physical sciences.[1–6] Generative foundation models, in particular, are displacing a swath of other methods. Foundation models involve extensive training of deep neural networks on enormous datasets in a task-agnostic manner.[7,8] Generative methods (genAI), often employing the transformer architecture,[9] seek to create novel outputs that conform to the statistical structure of training data,[10,11] enabling (*e.g.*) image synthesis[12–14] or text generation.[15] Large language models (LLMs) are generative models trained on text completion, but which can be adapted to a variety of tasks, including text classification, sentiment analysis, code or document generation, or interactive chatbots that respond to users in natural language.[7,16–18] The performance of LLMs increases with the scale of the training data, network size, and training time.[19–21] There is growing evidence that LLMs do not merely reproduce surface statistics, but learn a meaningful world model;[22–27] one correspondingly observes sudden leaps in capabilities during training, suggesting the emergent learning of generalized concepts.[25,28–31] LLMs can be tailored *via* reinforcement learning using human feedback (RLHF),[32–35] so that particular behaviors (*e.g.* helpful and truthful) are emphasized during generation. Generation quality can be improved by connecting to a corpus of trusted documents, which allows production of replies that are sourced and grounded (so-called retrieval augmented generation, RAG).[36–39]

While LLMs are often viewed purely as text-generators (*e.g.* for chat interactions), they have transformative potential owing to their ability to generate decisions and plans. For instance, LLMs can trigger software tools by providing them access to application programming interfaces (APIs).[42–50] Generations can be improved by inducing self-critique of output quality,[51–53] or creating chains of thought through iterative self-prompting.[54–56] These systems can be turned into task-oriented autonomous agents by allowing them to iteratively propose and execute solutions.[46,57–60]

The impressive capabilities of LLMs presage a paradigm shift in the way intellectual work is performed, as they empower humans to delegate many tasks to the LLM and instead focus on the highest-level deliberation and planning. However, there remain many outstanding questions about what system architecture and human–computer interactions (HCI) will best leverage these capabilities. Adaptation of these methods to scientific domains requires even deeper consideration, as science and engineering tasks are extremely technical and require high reliability and sourcing for both information and arguments.

Here, we explore the concept of an exocortex—an artificial extension to the human brain that provides additional cognitive capabilities. While future implementations of this concept might employ brain–computer interfaces (BCIs),[61] we argue that progress can be made by leveraging existing HCI methods to connect the human to a swarm of inter-communicating AI agents. If the individual agents are sufficiently capable, and their interactions sufficiently coherent, then the emergent activity could feel, to the human operator, as an empowering expansion to their mental capabilities.

We focus in particular on the concept of a science exocortex—meant to expand a researcher's intelligence and scientific reasoning—and propose some concrete architectural ideas. We propose an implementation (Fig. 1) using a swarm of AI agents that operate on behalf of the human user, and which—importantly—communicate with one another and thereby reserve human interaction only for high-value ideas and

Center for Functional Nanomaterials, Brookhaven National Laboratory, Upton, New York 11973, USA. E-mail: kyager@bnl.gov

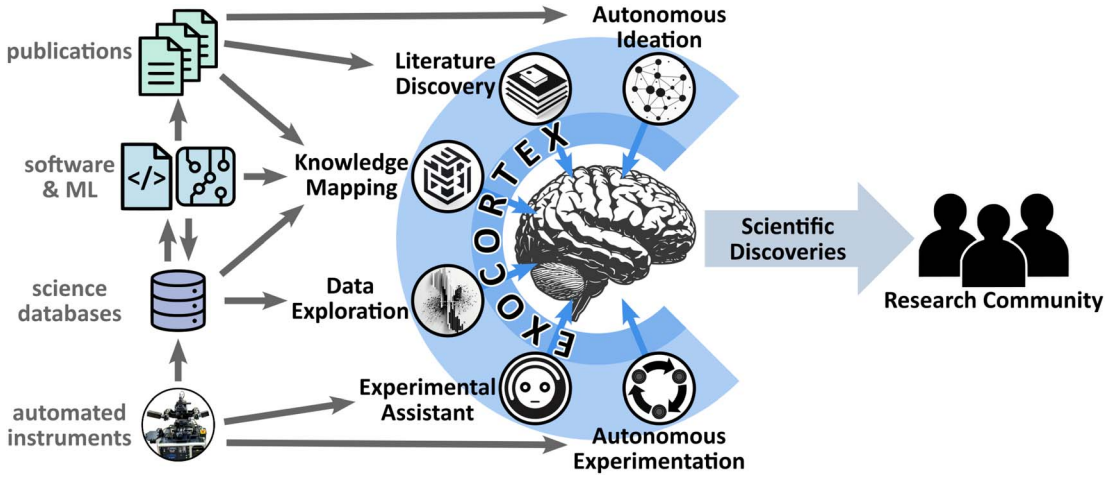

Fig. 1 An exocortex seeks to augment human intelligence by connecting computation systems to a person. A science exocortex could be implemented as a swarm of specialized AI agents, operating on behalf of the human researcher, including agents for controlling experimental systems, for exploring data and synthesizing it into knowledge, and for exploring literature and ideation. The AI agents would connect to science components (instruments, databases, software, *etc.*) and streamline access. Crucially, the AI agents communicate with one another, working on tasks on behalf of the user and only surfacing the most important decisions and outputs for human consideration. If successful, such a system would allow researchers to handle the enormity of modern scientific knowledge, and accelerate discovery and dissemination of new science.

important decisions. We define specific categories of required agents, including some focused on orchestrating experiments, others on data and software, and others on scientific literature. Although highly speculative, we hope the ideas presented herein stimulate further research on AI agents optimized for science, and their integration into systems that empower human researchers.

2 Discussion

LLMs natively output streams of tokens, and are by default used to generate text for humans to read, as in the canonical use as chatbots. However, a narrow interpretation of LLMs would miss their most significant capability: their outputs can be used as decisions, allowing one to automate (simple) cognitive tasks. Karpathy provides a provocative vision for the future of LLMs, wherein they act as kernels (orchestration agents) of a diverse set of capabilities (Fig. 2).[40,41] In a conventional operating system (OS), the kernel is a privileged software process that manages resource distribution and inter-process communication, allowing the end-user to access software systems, files on disk, network resources, and other services. By analogy, one can imagine a sort of AI OS, where the orchestration abilities of the LLM are leveraged to intelligently trigger the appropriate tool (*via* APIs,[42–50] code execution, *etc.*), retrieve relevant content (*via* RAG,[36–39] web browsing, *etc.*), and reformulate it into a form suitable for human consumption (text, images, audio, *etc.*). The crucial insight is that the LLM enables orchestration of tasks and resources, and aggregation of data sources, in a much more abstracted and high-level manner than is traditionally thought of as possible for software systems.

The exocortex concept takes this idea seriously, and expands upon it to propose that a swarm of agents could handle complex tasks. Each agent would operate in the manner depicted in Fig. 2, optimized for a particular task (by tailoring the available tools/documents, the prompting and scaffolding that dictate its input/output behavior, *etc.*). The interaction of agents, each acting as a sort of primitive cognitive module, could then lead to emergent capabilities in the whole.

Achieving this vision will be difficult, requiring solving a cascade of research challenges. Research is required to determine how best to exploit LLMs to generate agentic modules that can perform tasks autonomously (over short timescales) by iterating on a problem. Specialization of these agents to scientific problems will require additional consideration. The software infrastructure to have agents run over longer time periods, and inter-communicate productively, will

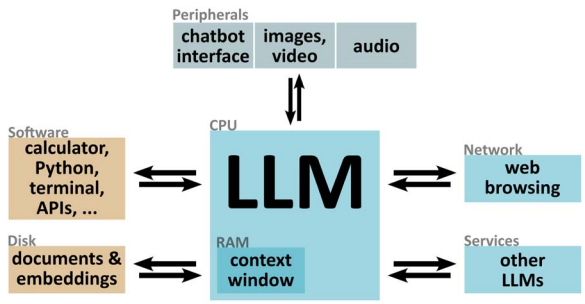

Fig. 2 Diagram of a large language model (LLM) acting as a kernel (image based on social media post by Andrej Karpathy). While LLMs perform text generation, Karpathy has proposed to view them instead as kernels—orchestration agents—of a new kind of operating system.[40,41] In this paradigm, the LLM is responsible for accessing resources (*e.g.* documents) or triggering actions (calculations, web browsing, *etc.*), and feeding results to a desired interface (*e.g.* chatbot dialog). The ability of LLMs to perform (rudimentary) decision-making can thus be exploited to coordinate more complex activity in response to relatively vague commands (which may come from a human or another LLM system).

need to be developed. The correct inter-agent organizational and communication structure will need to be identified. And, finally, the appropriate interface between the ecosystem of AI agents, and the human operator, will need to be developed. Below we provide initial thoughts on these various challenges.

2.1 AI agents

Research into LLM-based AI agents is ongoing, with several prototypes having been demonstrated.[46,57–59,62–66] Although the optimal architecture remains an open question, current demos typically add several elements to the base LLM, such as: providing the LLM with access to various tools (software, web browsing, *etc.*), the ability to store information about its ongoing work (*i.e.* memory[67–70]), some kind of loop to iterate on problems (inner monologue[71] or chain-of-thought[52,54–56] or internal graph search[72]), and prompting suggesting breaking a problem into steps, and then working progressively on each step. Further improvements are possible using an architecture where task plans and status are captured in an explicit tree structure, which provides a flexible way to organize complex hierarchies.[60] Tree structures can be efficiently searched (*e.g.* Monte Carlo tree search) for reasoning and planning, yielding improvements in many tasks including math.[73–77]

Additional research will be required to adapt agentic LLM approaches to scientific problems. Straightforward improvements would arise from training or fine-tuning LLMs on scientific documents, to ensure understanding of the relevant topics. Fine-tuning on math examples can elicit latent mathematical abilities.[78,79] Document retrieval can also easily improve LLM performance on scientific tasks.[37] Additional LLM specializations for science should also be considered. Golkar *et al.* proposed xVal, a specialized token encoding for numbers (scaling a dedicated embedding vector) which improves LLM handling of numerical tasks.[80] McLeish *et al.* used special positional embeddings (relative to start of number) and demonstrated vastly improved performance and generalization on simple arithmetic (addition and multiplication) tasks.[81] Xu *et al.* integrated symbolic expressions and logic rules into a chain-of-thought prompting strategy, demonstrating improved reasoning on logical tasks since the LLM was invoking formal logic and symbol manipulation during the solution.[56] Trinh *et al.* combined a language model with a symbolic solver to handle geometry theorems.[82] Vashishtha *et al.* improved causal reasoning by providing axiomatic training examples.[83] These kinds of approaches appear promising, suggesting that LLMs with slight adaptations could yield vastly improved reasoning for science and engineering tasks.

An advantage of the exocortex architecture is that it can easily integrate more advanced AI agents as they are developed by others. In other words, we propose to separate the design/function of agents from their inter-communication, so that new agents can be added to the exocortex easily (by simply building a wrapper that supports the expected messaging between agents). The goal is to be able to leverage the growing ecosystem of AI modules being developed for science, including for simulating complex systems,[84] optimizing differential equations,[85] fluid dynamics,[86] material discovery,[87] crystallography,[88] 2D materials,[89] chemistry tools,[64,90] protein representation[91] and design,[92] and pathology images.[93,94]

2.1.1 Autonomous experimentation. Autonomous experimentation (AE) is an emerging paradigm for accelerating scientific discovery, leveraging AI/ML to automate the entire experimental loop, notably the decision-making step.[95,96] AE aims not to merely quantitatively accelerate, but also to qualitatively improve experiment execution by having an algorithm adaptively select optimal experiments. It seeks not to replace the human researcher, but to liberate them to operate at a higher level of abstraction where they can focus on scientific meaning instead of micro-managing experimental details.[96,97] Progress in AE has grown rapidly over the last few years, transitioning from proof-of-principle to true discovery of new science.[98–116] The AI control module may exploit reinforcement learning,[110,117] though a highly popular approach is to exploit Bayesian methods[118] (such as a Gaussian process,[107,119] GP) since this provides rigorous modeling of a data surrogate and associated uncertainty. AE methods are increasing in sophistication, including demonstration of multi-modal autonomous experiments integrating multiple measurement systems.[120]

Instead of treating the AE system as an autonomous loop initiated and monitored by the human researcher, one can envision it as a module in the exocortex, which can be activated and monitored by other AI agents. Enabling this capability would require relatively little change to existing AE architectures. Primarily, one would need to define a simple software API or natural-language interface for AE parameters and actions. Doing so would increase the power of AE systems, as they could more easily integrate physics-informed priors arising from literature or preexisting datasets.

2.1.2 Experimental assistant. A highly consequential type of AI agent for science is one that negotiates control of some experimental tool on behalf of the researcher. Building such an agent requires the experimental system to already be highly automated, such that the agent can trigger operations (synthesis, measurement, *etc.*) and retrieve generated data. However, as more and more platforms naturally shift towards higher levels of automation, the prospects for AI control improve. Many high-end measurement tools provide highly software-driven interfaces, including electron microscopes,[108,109] scanning probe instruments,[108,121,122] and synchrotron[114,123] or free electron laser (FEL)[124–126] beamlines. Recent work has also demonstrated automated workflows[115,127–130] or modular platforms[131–133] for lab experiments. The rapidly advancing capabilities of AI robotic control[134–140] suggest that broader ranges of manual laboratory tasks will soon be amenable to automation.

The next step is thus to design AI agents that can access the capabilities of these automated systems. Preliminary work has already demonstrated the viability of LLM-based agents for controlling scientific instruments.[90,141–143] AI experimental assistants can allow the human to phrase commands in natural-language, whereupon the LLM can convert this into action through API calls or code generation.[5] The assistant can help to integrate experimental and data analysis steps, making it easier

to see the consequences of measurements and to iterate more quickly on the problem being studied. Experimental assistants can also act as tutors, *e.g.* generating initial control code for a user unfamiliar with a particular instrument. The approach is flexible, and can easily be adapted to changing instrument conditions by updating documentation that is added to the LLM's context during operation.

2.1.3 Data exploration. Scientific discovery involves collecting, processing, and analyzing datasets of many types. In answering a scientific problem, researchers will integrate a wide variety of data sources, including lab notes, instrument outputs (images, spectra, *etc.*), simulation results, and a succession of derivative data products created through analysis. Tracking, organizing, and visualizing these datasets is extremely challenging. An acute challenge for a modern researcher is the interdisciplinary nature of many frontier topics, which correspondingly means dealing with a heterogeneity of datasets coming from different sources, and following different conventions for formatting and meta-data.

AI assistants could play an important role in alleviating this burden, by automating many routine tasks in data triage and reformatting, and by automatically triggering the required pipeline for automated data processing. The heterogeneity of data can possibly be handled using foundation models.[8] Whereas in the past, application of machine-learning to science required training bespoke models on carefully-labelled datasets specific that science topic, foundation models trained on vast quantities of unlabelled data should be able to learn generic representations that are useful. For instance, it was found that the Contrastive Language-Image Pre-training (CLIP) model[144] trained on generic Internet image data could be used to assess similarity for scanning electron microscopy and X-ray scattering datasets, without any retraining or fine-tuning.[37]

Many scientific datasets are images, or can be converted into images. Thus a powerful approach for AI data assistants is to exploit multi-modal language/vision models.[94,144–147] Indeed, humans generally consume data as images, in the form of graphs and plots; this visual formulation of the data has the advantage of being ready-to-deploy, human-readable, and already well-represented in existing training sources (publications). Exploiting multi-modal models as data assistants is still in its infancy, but early systems[93,148–152] show promise.

2.1.4 Knowledge mapping. A grand challenge in data science is to integrate data from disparate sources into a single model. Human scientists excel at this task, as they integrate insights provided from experimental data, calculations, literature they recall, and intuitions informed by years of scientific practice. When thinking about or discussing a complex topic, human scientists will naturally jump between different levels of abstraction and thus different scientific models. This combination of models (some highly quantitative, others heuristic) allows human scientists to compensate for the deficiencies in one model/reasoning by leveraging another. Approximating this efficient behavior in a synthetic knowledge system is challenging. AI analogs of human synthesis would make knowledge integration explicit and documented, and provide integrated models for other software systems to leverage.

A core challenge is to align disparate observations of the same physical signal; that is, to account for the unknown disparity between models arising from systematic errors, different underlying assumptions, mismatched definitions, *etc.* Consider a simple signal for a material system (*e.g.* crystalline grain size) as a function of a physical parameter (*e.g.* temperature). Despite the independent/dependent variables being well-defined within one model (observation), matching between models may not be trivial. For instance, a physical measurement might use absolute real-world temperature (in kelvin units) while an associated coarse-grained simulation might rely on a unitless abstracted temperature variable. (Obviously the two quantities are closely related; but the mapping function between them is typically not known.) The measurement of grain size by different techniques may not match owing to different definitions (*e.g.* volumetric *vs.* aerial averaging).[153] Thus, it can be challenging to merge datasets relevant to the same physical problem, even when they are individually trustworthy and robust. The problem becomes harder still as input data sources become more heterogeneous and ill-defined (heuristic classifications, text descriptions, scientist intuitions, *etc.*).

The simplest approach to this problem might be to exploit contrastive learning. For instance, the CLIP[144] model uses two encoder pathways: one for text and one for images. The method also computes a similarity matrix between the two latent spaces (cosine similarity for text/image pairs), where part of the training loss seeks to maximize the diagonal and minimize the off-diagonal elements. In this way, the text and image latent spaces align, allowing cross-modal learning and applications. In principle, a similar approach could be used for scientific data. Datasets and their associated text descriptions could be used as training pairs, or different observations of the same physical phenomenon (*e.g.* experimental measurements and corresponding simulations) could be combined if some pairwise associations were manually identified. Recent work increasing the number of modalities appears promising.[154,155] Nevertheless, it may be challenging to scale this approach to handle the heterogeneity, complexity, and sparsity of realistic laboratory datasets.

A more sophisticated approach to this problem is to train multi-modal foundation models on scientific datasets.[6] The Polymathic AI effort is proposing to train AI models for science on a breadth of data,[156] which can then be specialized for any particular application by exploiting the latent representations or *via* problem-specific fine-tuning. Cranmer argues that doing otherwise (*e.g.* training an ML model for science using random initialization) is inefficient as it ignores the wealth of well-understood scientific priors.[157] Initial results for this approach are promising,[158,159] with (*e.g.*) multi-physics pretraining on system dynamics improving subsequent predictions on new systems. The approach involves projecting the fields for different kinds of physical systems into a shared embedding space. The central generative model (based on transformers) thus learns meaningful physics, while the dataset-specific embedding/normalization schemes capture the differences between the physical systems.

A closely-related approach would be to train multi-modal foundation models on science data, so that these models could be queried to explore trends in the data. Recent work[160] has shown that an LLM trained on (x, y) pairs can articulate the function $f(x)$ that underlies the transformation (can define it in code, can invert it, *etc.*). If this result generalizes, it implies that LLMs trained on raw science data could coherently describe the data, make predictions based on the underlying functions, and so on.

A different way to formulate this task (Fig. 3) is to focus on integrated modeling of all the signals defined in physical parameter spaces for a given problem (*e.g.* class of materials). For a given signal, a variety of different observations might be available (from experiments, simulation, theory, *etc.*), with tradeoffs between signals (in terms of sampling density, error bars, validity in different parts of parameter space, *etc.*). Signals could be combined into a merged model by learning a non-linear transformation that maps them into a common space and maximizes their overlap (using, *e.g.*, variants of the methods described above). The combined datasets could be interpolated using a Gaussian process or other nonparametric method,[107,119] leveraging physics-informed constraints to further improve the model (*e.g. via* tailored kernel design[113,161]) and acceleration methods to reduce computation cost.[162,163] The set of models could also be cross-correlated, to identify connections and scientific trends between signals (or establishing lack thereof). GP modeling of correlated signals would also allow interpolation of signals into parts of spaces where they were not explicitly measured (effectively using a correlated measurement as a sort of proxy signal). Conceptually, the set of signals (and covariance matrix between them) represents a final rich multi-modal model of full system behavior. This unified model could be used for predictions, searching for trends and novel physics, or as a guide for future discovery (identifying under-sampled regions, suggesting high-performing materials, *etc.*).

An even more speculative approach would be to attempt to adapt methods of generative world synthesis to scientific data. There has been enormous progress in generative synthesis of images (2D data),[12–14] objects (3D),[164–173] and video (3D).[174–179] Neural radiance fields[180] and Gaussian splatting[181] have emerged as efficient methods for reconstructing and representing 3D scenes (where input images act as projective constraints). These methods have been extended to capture[182–184] or synthesize[185–189] changes over time (4D). These methods are efficient,[190,191] scalable,[192] and amenable to in/out-painting.[193,194] In addition to obvious applications in content generation, these methods are seeing adoption for autonomous

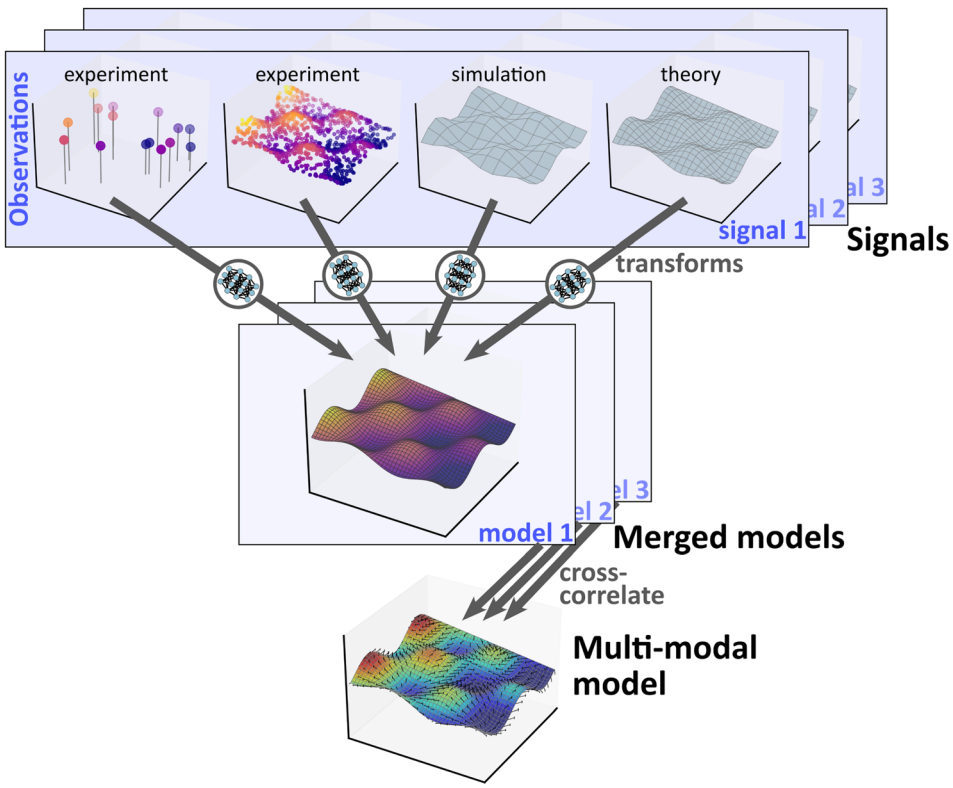

Fig. 3 Knowledge mapping is an attempt to align and aggregate a variety of data sources about a particular scientific problem into a single model. One architecture for accomplishing this is shown. Available data is organized into signals of interest (such as physical measurables, material properties, or functional metrics). One typically has a variety of estimates or observations for a given signal, arising from different experiments, calculations, or theories. In principle these observations already map into a common space; in practice there are complex and often unknown disparities between the observations, owing to measurement errors, disparate definitions, or different assumptions. Thus, some non-linear transformation (*e.g.* accomplished using neural networks) is required to combine them into a single predictive model. Models for distinct signals can be cross-correlated to identify inter-relations; this can effectively combine the models into a single multi-modal model.

driving[195] and robotics.[196–199] The trajectory of these development suggests neural synthesis of virtual world,[179] wherein immersive 3D environments are generated and animated/evolved, using real-world reconstructions and/or user text commands as inputs. We suggest that this approach could be applied, in higher-dimensional spaces, to scientific datasets. The partial measurements made in scientific experiments can act as constraints (conceptually a projective view of the full higher-dimensional space), where the objective is to reconstruct a consolidated model consistent with all the data (*i.e.* merge datasets and modalities) and to generatively fill unmeasured parts of the space using an informed model (*i.e.* interpolate and extrapolate in a physics-aware manner). Considerable work would be required to recast existing methods to handle the dimensionality and different constraints of scientific data in physical parameter spaces; but the efficient representations being developed for simulating the real world may well hold useful insights for representing other kinds of coherent data-spaces.

2.1.5 Literature discovery. The scale of the research literature is continually growing, making it increasingly difficult for researchers to maintain awareness of important trends or singular results. Conversely, this enormous scientific corpus is a trove of insights that should be more fully leveraged. Literature Based Discovery (LBD)[200–202] has a long history of increasingly sophisticated methods and software being developed for mining the literature, identifying connections across domains, and otherwise streamlining literature research. AI methods, and LLMs in particular, are well-positioned to greatly accelerate these processes, automating knowledge extraction from publications.[203]

An obvious use-case is to systematically search through a corpus in order to extract and tabulate values for quantities of interest.[204–206] Here, the flexibility of LLMs can enable extraction that handles the heterogeneity arising from synonyms, different definitions or units of measure, and so on.

LLMs can be combined with document retrieval (RAG) to allow users to rapidly identify relevant documents (or subsections thereof) and immediately incorporate them into reasoning or question-answering. RAG LLMs have been used to build domain-specific chatbots for science,[37] and to provide an interface to vast materials data that can be distilled as requested by the user.[206] More generally, the AI model can be exploited as a co-pilot to help the user access specialized knowledge or tools, as has been demonstrated for catalyst research,[207] chemistry experiments,[59] and chemistry tools.[90] A valid concern is that the reasoning of LLMs—being well below human scientists—would be insufficient to be useful. However, when designed as a co-pilot, such systems can offer substantial value. LLMs can exploit the systematic compositionality of language (and thus ideas), which enables them to generalize in useful ways.[208] Evidence shows that dialoging with LLMs can indeed help researchers.[2,209]

LLMs can also be exploited to automate tedious tasks. For instance, they can be used for ranking,[37,210,211] evaluating,[212] or classifying[37] scientific documents. This opens up a new possibility for researcher engagement with the literature, beyond the conventional activities of periodically searching for articles of interest and keeping a watch for relevant articles through networks (peers or automated). LLMs could be used to search, organize, rank, triage, and summarize papers, and thereby identify the most pertinent publications for human consideration.

LBD has a strong history of exploiting network analysis to understand the science corpus, including using predictive knowledge networks.[213,214] An interesting possibility would be to exploit modern foundation models as another form of network analysis. The semantic embedding provided by these models could offer a rich means of identifying connections (or lack thereof) in the literature. For instance, clusters of publications that are semantically similar but not cross-citing one other could represent inefficiency (duplicative efforts unaware of each other), while clusters that are highly correlated in a subset of embedding dimensions (but divergent in others) could represent opportunities for collaboration.

Another use for AI agents is to aid researchers in drafting scientific manuscripts. LLMs are fundamentally text-generation systems, and their role in productively generating long-form textual content is being extensively studied.[215–217] As often observed with LLMs, the quality of output can be improved through iteration, including using the LLM to generate an outline, self-critique output, and so on. There are additional challenges in using LLMs to generate scientific text, as consistency and correctness must not be compromised. Here too, there has been progress in using LLMs to automatically generate full-length technical documents.[218–220] The use of AI to generate text for inclusion in the scientific literature could be deleterious to science if the texts contain too many errors (compared to the human baseline). The exocortex design emphasizes the central role of the human researcher in assessing correctness and validating decisions/generations. We propose that the human researcher maintain the important role of validation, and thereby maintain responsibility for the quality of publications to which they attach their name.

2.1.6 Autonomous ideation. A novel use for LLMs would be to help automate the task of generating and evaluating scientific ideas, including research plans, testable hypotheses, experimental plans, and predictive theories. These cognitive tasks are among the most high-level performed by human scientists, and as such least likely to be fully automated by LLMs in the foreseeable future. On the other hand, the process of human ideation involves many secondary cognitive activities that could be automated.[221] Thus, autonomous ideation seeks to generate loops of machine-driven brainstorming and evaluation, bringing high-value ideas to the human's attention for further consideration.

Existing work in LBD has begun to tackle the question of how to use natural language processing and LLMs for hypothesis generation or other scientific ideation tasks. A central question is whether LLMs can be creative at all. LLMs are trained statistically on a large document corpus, and can be viewed as generating novel text that are interpolations in a semantic space. Such generations can be factual (correctly composing ideas in the training data) or erroneous "hallucinations" (or

confabulations). Hallucinations can be partially mitigated by detecting them through generation uncertainty,[222] or by grounding responses using RAG.[36–39] Although hallucinations are generally undesirable, their existence is intrinsic[223] and there is a tradeoff between hallucinations and creativity.[224] In other words, some amount of hallucination is desirable, to enhance creativity and communication.[225] More broadly, evaluations of LLM creativity suggest that they can generate outputs that are non-trivially novel and useful to humans.[226–230] Language models have demonstrated utility for hypothesis generation,[231,232] or as generators for novel ideas.[221,233,234]

The most direct way to use LLMs for ideation is as a chatbot assistant to a human researcher. A more automated design would leverage agentic AI operating in loops, so that a group of LLMs propose and critique ideas, and then rank[37,210] these ideas in order to identify the most promising. A more structured (but speculative) approach is to treat the task of autonomous ideation as being analogous to autonomous experimentation,[96] wherein an ML decision-making algorithm selects points in a physical parameter space for measurement. In autonomous ideation, one could analogously select points in the semantic "space of ideas" for exploration (Fig. 4). More specifically, the search space is defined using a semantic vector (*e.g.* text embedding) and the target signal in that space is defined using LLM ranking of the ideas. On each loop, a new region is selected for exploration, using a modeling process that can consider both idea ranking (bias towards high-quality regions), and uncertainty (explore under-sampled or high-error regions). This modeling can exploit Gaussian process methods to naturally capture uncertainty and learn hyper-parameters that describe the semantics being explored. Once a point is selected, an LLM generates new ideas at that position by (*e.g.*) sampling a local neighborhood[234] of ideas or documents in order to generate new content. This generation is ranked to quantify it as a signal, which is fed back into the loop. As this procedure continues, it will naturally fill the semantic space of ideas, balancing between exploration and exploitation, and providing a surrogate model for idea quality in the subspace selected for search. It is an open question whether Bayesian modeling can meaningfully be applied to the inherently vague space and signals associated with ideation. But the AE framework provides a robust starting point for rigorously testing various idea exploration schemes.

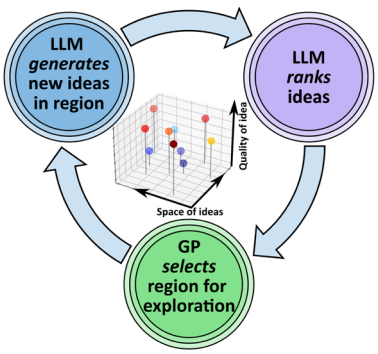

Fig. 4 Autonomous ideation aims for the AI agent to develop new scientific ideas (novel research directions, testable hypotheses, actionable research plans). One possible system design is to treat the task similar to an autonomous experimentation loop, wherein one is exploring a multi-dimensional parameter space. In ideation, one can define the space of ideas using embedding vectors to position each idea. Each idea can be scored using an LLM ranking procedure. The loop consists of selecting a region for exploration (*e.g.* based on some combination of local sparsity, model error, and quality-maximization), generating ideas in that region (*e.g.* using an LLM provided with documents/ideas from the local neighborhood), and ranking the resultant ideas. As the loop proceeds, the space becomes populated with ideas. The top generations can eventually be presented to the human for consideration.

A different ideation design would be to leverage ongoing work in visualizing and interpreting the internal state of the LLM. While neural networks are often described as inscrutable black boxes, there has been enormous progress in interpreting their structure and the latent spaces in which they operate. In vision models, the role of neurons and circuits can be interpreted by visualizing strong activation patterns.[235] In language models, tasks learned in-context can be understood as a simple function vector that capture the relevant input–output behavior.[236,237] A particular direction in the model's internal state can be associated with specific behavior, such as refusal to respond[238] (allowing that behavior to be selectively amplified or weakened). Identifying internal circuits associated with particular concepts allows one to build "circuit breakers" to suppress undesired output.[239] Natural hierarchies of concepts—which occur throughout natural language and especially in scientific ontologies—are represented in the model's internal vectorial space as polytopes that can be decomposed into simplexes of mutually-exclusive categories.[240,241] Model activations can be interpreted using human concepts, if they are projected into a higher-dimensional space to disentangle them.[242–244] These interpretability insights are often exploited for alignment,[245] to elicit safe and desirable model behavior. However, they could also be used to directly explore the landscape of ideas. For instance, visualizing the internal ontology for a scientific sub-space might allow researchers to identify regions of unexplored concepts, or to see fruitful cross-connections between ideas that are typically considered unrelated. Searching the structure of this space for common patterns could further reveal new connections or universal motifs. Being able to directly alter the activation or geometry of the semantic space, and observing LLM output, provides another avenue for generating novel ideas in a highly directed way. This research thrust would require researchers building new intuitions about how to understand and navigate the complex spaces internal to LLMs.

More generally, strategies originally intended for model tuning or alignment could all be co-opted for ideation. For instance, one could block off exploration of ideas known to be fruitless, or conversely emphasize desired modes-of-thought. Viable strategies include fine-tuning,[246–248] RLHF[32–34] (including AI-assisted[35]), constitutional adherence,[249] preference ranking,[250] instruction backtranslation,[251] principle-driven self-alignment,[252] or eliciting latent knowledge.[253–255]

2.2 Exocortex system

The proposed exocortex design will behave as a system of interconnected AI agents, some of which can also communicate

directly with the human researcher. The correct design for this system is an open research and engineering challenge. Nevertheless, we can begin to propose and test designs. One of the simplest implementations would be for each researcher to build a personalized network by selecting among pre-existing AI agents, and defining connections between them based on desired workflows. Communication between agents could thus be managed with point-to-point message queues. This approach is not very scalable, however. An alternative would be to establish a central database where inter-agent messages are accumulated, and build code that manages communications, using user-defined heuristics to decide when incoming messages require returning to the same agent for revision, launching a new agent, passing to a running agent, or bringing to the human's attention. Likely there are yet better designs possible if one treats the agent-interaction problem as a large machine-learning task. By selecting a flexible design (*e.g.* based on graph neural networks), an automated optimization process could create/eliminate connections in order to build dynamic workflows. Much of the work on within-agent iteration and looping can be exploited to improve inter-agent workflows.

In all these schemes, signals between agents can take the form of plaintext messages. This has the advantage of being highly legible to the human operators,[256] allowing them to understand commands, make improvements, and even extract scientific value from intermediate products. As the number of agent types increases, the diversity of possible inter-agent cooperations increases quadratically, while the space of possible workflows grows exponentially. Example messages that might be sent between agents are shown in Table 1. Legible inter-agent messages will allow the human operator to inspect, at will, operation of the system, including editing an agent's message before it is executed by another agent.

The complexity of interconnected agents, and the non-standardized (text-based) messaging between them poses a problem for automated monitoring, analysis, and optimization of these systems. On the other hand, it is possible that existing approaches for systems engineering can be recast to the context of AI swarms. For instance, machine-learning has benefiting enormously from gradient backpropagation,[257] which has essentially automated the process of optimizing complex neural network and AI models. By analogy, Yuksekgonul *et al.* proposed TextGrad as a text-based "differentiation" of AI systems.[258] Natural language feedback (*e.g.* criticism) of system outputs can be used as scores (analogous to loss), the variation in score as a function of changes in prompt can be used as a gradient, and gradients can be propagated across the system with knowledge of architecture. This allows automated optimization of LLM-interaction networks. Zhou *et al.* demonstrate how symbolic learning can be applied to optimizing LLM frameworks.[259] Further developing techniques such as these may be crucial to properly optimizing exocortex-like systems.

The proposed exocortex architecture would treat agents as modules, allowing them to be swapped or for new agents to be added. Connecting agents to each other should enable progressively more complex automated workflows. However, a crucial open question is whether multi-agent workflows can scale to complex problems. For instance, even with a low per-step error rate, long task sequences could easily accumulate intolerable total error rates. The acceptable error-rate will be quite different for different parts of a workflow. For instance, imperfect ideas generated during ideation have low risk, as they will be identified and filtered out by the human easily (in human ideation there is value in initially considering erroneous ideas, as this can improve creativity). On the other hand, errors introduced by a data-analysis agent could be subtle and difficult to detect; yet errors in this stage would contaminate downstream analysis and thus invalidate the science. Errors in the experimental stage could waste valuable resources (time, experimental material, *etc.*) but are likely to be caught by human oversight. AI agent workflows can also be difficult to debug (owing to stochastic response) and brittle to maintain (changes in cloud models, changes in input data distribution, *etc.*). Thus, the scalability of multi-agent workflows is a crucial open question, requiring research and development. Frontier work in this area suggests that well-designed multi-step AI workflows may be able to generate coherent outputs.[219]

The goal of the exocortex is to augment a human scientist's intelligence. This objective is predicated on the assumption of emergence at two levels: one, that the swarm of AI agents will, through coordination, exhibit intelligence greater than the naive summation of their respective abilities; and two, that the combination of exocortex agents and human thinking will enable greater effective intelligence. To succeed, the exocortex architecture must thus enable this outcome. The correct design remains an open research question. However, we propose that analogies to human cognition can aid in the design.

2.2.1 AI–AI interactions. LLMs generate ideas and decisions, but they are quite primitive in the sense that the ideas are reflexive rather than resulting from deep introspection.[256] The repeated waves of processing that occur within an LLM as it proceeds through tokens provides an opportunity to build-up more complex assessments, with the current understanding represented as updates to the residual stream. Improved behavior can thus be elicited by inducing the model to explicitly output reasoning steps.[260,261] Interestingly, introducing even meaningless filler tokens into the output provides improved performance,[262] presumably owing to the additional computation cycles that are invoked. And yet, LLMs implement a relatively primitive and unidirectional method of thinking, as they are unable to revise the serialized output. Multiple research efforts aim to improve this by introducing a sort of deliberation cycle, such as by triggering self-critique of output,[51,52] or generating chains of thought through iterative self-prompting.[54–56] Exploiting tree search (*e.g.* Monte Carlo) can further improve quality, especially on math problems.[73–76] For scientific applications, versions of these methods that explicitly invoke formal logic are especially attractive.[56,83] One can also provide pre-designed thought-templates to improve reasoning on selected tasks;[263] building a catalog of templates for scientific tasks would be beneficial.

Another means of generating improved output is to construct "societies" of semi-specialized AI agents, and allow them to communicate and cooperate on a task. The hope is that

Table 1 Examples of command messages that various AI agents could send to other agents. The diagonal elements (grey text) are commands sent from an agent to another instance of the same type

Message to ↓ / Message from →	Autonomous experimentation	Experimental assistant	Data exploration	Knowledge mapping	Literature discovery	Autonomous ideation
Autonomous experimentation	Relaunch this AE with tweaked parameters...	Launch new AE with these parameters...	Incorporate this data as AE prior...	Build AE kernel that conforms to this model...	Set AE parameters based on this literature...	Launch AE that tests this idea...
Experimental assistant	Queue these follow-up experiments...	Launch this sub-experiment...	Analyze these related datasets...	Overlay this model with current experiment...	Create plan that replicates this published experiment...	Prepare an experiment plan to test this idea...
Data exploration	Plot ongoing AE...	Retrieve and analyze data for this experiment...	Compare/integrate with this dataset...	Overlay this model on dataset...	Add citations to dataset...	Add annotations to dataset...
Knowledge mapping	Test integration of AE into model...	Add this experiment to model...	Integrate this data into model...	Compare this model with past models...	Integrate literature results into model...	Annotate model with this interpretation...
Literature discovery	Is this trend expected, based on prior work...	Search literature for the value of...	Find data values relevant to...	Find models/equations relevant to...	Find and summarize papers on this sub-topic...	Retrieve literature relevant to this topic-space...
Autonomous ideation	Evaluate progress of this ongoing AE...	Generate hypotheses relevant to current experiment...	Generate hypotheses for this dataset...	Propose theories to explain this model...	Launch ideation based on these papers...	Launch ideation on this sub-topic...

specialization improves diversity and allows task-specific targeting, and that the emergent quality of collective output is higher than for any individual agent. Although this approach is only nascent,[58,264–270] there are early suggestions that it can improve task performance in contrived contexts (*e.g.* games[271,272]) and applications (*e.g.* code generation[273] and translation[274]). One can also use synthetic analogs of cultural transmission to improve learning of AI swarms.[275,276] Interaction between agents can be[269] cooperative, debating, or competitive. Agents can be organized into flat structures, where each agent is equivalent (*e.g.* voting on answers/decisions), or hierarchically, where top-level agents assign tasks to workers, and aggregate outputs. Different tasks will, of course, call for different organizational structures. However, there are often clear advantages to establishing hierarchies and workflows,[267] especially where one can draw inspiration from human organizational structures.

Instead of emulating human social structures, an alternate architecture is mixture-of-agents,[270] which organizes AI blocks into layers reminiscent of neural networks (where each node is an LLM instead of a synapse). The input prompt is fed into a layer of models that propose independent responses, an aggregator synthesizes the responses into an improved output, and this is fed into the next layer. Thus, response quality progressively improves across layers, as more reconsideration is performed. By including different LLMs within a layer, one can improve diversity and allow for models to compensate for each other's weaknesses. Performance can also be optimized by correct selection of models within layers. The architecture is rationalized and organized, and amenable to rescaling (changing number of agents per layer, number of layers, *etc.*) to optimize for a particular task. This work demonstrated significantly improved outputs, compared to single-shot use of any underlying model, and demonstrated that a final aggregation LLM call (rather than ranking and selecting the best output so far) improves generation. This supports the idea that agent interactions can lead to emergent capabilities greater than any individual agent. The iterative processing may also make multi-agent setups amenable to longer-horizon tasks (*e.g.* longer text analysis or generation).

The optimal architecture for providing LLMs with deliberative capabilities remains an open and exciting research question.[256] Current scaling suggests that LLMs have untapped potential that could be unlocked with appropriate designs. In parallel with algorithmic research, we propose that scientific researchers can make progress by simply expending compute to compensate for architectural weaknesses. For instance, consider tasking an agent-swarm with a problem, whereupon the agents generate ideas, ask each other questions, generate random permutations by combining ideas, rank all ideas, and only present the best ideas to the human. This workflow is highly wasteful in the sense that the majority of the generated content is never seen and indeed low-quality. Yet this invisible content can be viewed as the system's internal deliberation. Even if this process is inefficient, its automated and unobtrusive nature can make the outputs sought-after by humans. In the context of science expenditures, the associated costs could be small relative to the value. There are tentative reports that this kind of extended search can yield substantial improvements.[277–283] We thus propose increasing investigation by physical scientists of brute-force workflows, for generating content useful to researchers.

2.2.2 Human–AI interactions. Human thinking involves a combination of effortless intuition and deliberative reasoning[284–287] (often referred to as "implicit" *vs.* "explicit" or as "system 1" *vs.* "system 2"). A cluster of low-level brain modules generate reflexive actions, intuitive assessments, and creative ideas. A higher-level deliberative process engages in discrimination, iterative refinement, and selection; using the low-level generators as inputs and assessors. A synthetic exocortex can be designed similarly. The swarm of AI agents act as low-level generators, introducing ideas and providing reflexive assessments. The human deliberative consciousness remains the core, doing the highest-level discrimination and decision-making, and is thus ultimately the locus of volition.

The exocortex interface should ideally make the AI-generated inputs feel much like the human's own low-level modules. When actively working on a task, the exocortex should provide contextual assessments and ideas that feel like spontaneous intuition that the human can trust (but will also verify). When returning to a dormant task, accumulated background AI-swarm processing should feel like the mental incubation known to occur in humans,[288] wherein returning to a problem after a diversion often yields new insights and perspective owing to subconscious consideration.

Obviously, efficient coupling between reflexive and deliberative processes is required in humans for effective creativity and problem-solving.[289] A legitimate concern is that traditional peripheral-based user interfaces (using keyboards, screens, *etc.*) represents too much friction for strong coupling, and brain–computer interfaces will be required.[61] However, there is ample evidence of human tool use becoming overlearned[290] to the point that the tool is considered an extension of the person's body and volition.[291] We can view the evolution of cognitive technology as precedent for humans externalizing aspects of their cognition, with a succession of tools (writing, calculators, the Internet, smartphones) being exploited as external memories, processing extensions, or task-activation schemes. Thus, we posit that fast and responsive interaction through existing computer interfaces may be sufficient for the desired interaction. Indeed, humans are known to be able to enter so-called "flow states" (immersed and focused)[292,293] during computer-oriented tasks such as programming.[294]

2.2.3 Human–computer interface. The purpose of the exocortex is to offer the human additional cognitive power that feels—as much as possible—as a natural extension to their own mind. One can imagine a future where brain–computer interfaces are used to provide and ideal interface;[61] we posit that in the short term much value can be realized by providing researchers with AI agents through traditional computer interfaces. Research in human use of autonomous tools suggest that the person must ultimately feel that they are in control of processes.[97] Correspondingly, we propose that initial exocortex interfaces will involve humans reviewing and verifying LLM plans before they are executed (by other AI agents). There is evidence that humans observing the output of LLMs debating each other helps the human identify the best ideas.[295,296] This suggests that, more generally, providing researchers with access to exocortex inter-communications (critique, debate, refinement, *etc.*) could provide them with valuable information. As system robustness improves, and user confidence in the tools increases, more and more workflows can be automated and unattended.

With respect to human interaction with the software tools, we can define several modalities:

- Push: where alerts are used to capture the user's attention (operating system notifications, text messages, *etc.*).
- Pull: which require the user to actively check on status (visiting web page, opening a program, *etc.*).
- Ambient: where information is displayed peripherally to the user, or where contextually relevant.

Different aspects of exocortex operation might imply a different notification mode. For instance, human-directed dialogue is inherently pull, while operationally-critical and time-sensitive statuses that require human resolution will be push. However, the ambient modality is the most well-aligned to the ethos of the exocortex, where information generated by AI agents is contextually but unobtrusively presented, available to subconscious consideration by the human, and thus appears to the user as a seamless extension of their ongoing planning.

In the short term, we can envision useful interfaces being developed by exploiting HCI best-practices for ambient information display, and by integrating exocortex outputs into existing visualization tools and workflows. Extended reality (virtual reality, augmented reality) tools may be natural peripherals for exocortex software. Leveraging improving systems for voice transcription and voice synthesis provides another avenue for natural interaction with these tools. We note that as LLMs increase in capability, they are beginning to develop a primitive theory of mind.[24,25,297,298] This can be taken advantage of by using the LLM to roughly model human behavior, and thereby providing suggestions in ways that are most beneficial and least disruptive.

2.3 Infrastructure

In addition to novel AI developments, the success of the exocortex requires continued progress in several pragmatic infrastructure components (left side of Fig. 1). In general, science infrastructure must be made increasingly automated and software-accessible, so that AI agents will be able to leverage these systems as tools. Importantly, even if the exocortex concept is flawed, the proposed improvements in science infrastructure will be of great value to the community.

2.3.1 Automated instruments. As previously discussed, scientific instruments are becoming increasingly automated. This trend is driven by the increasing complexity of these tools (there are too many layers of control for them all to be manually managed), and researcher desires for speed and efficiency. Automated tools are in principle amenable for activation by AI agents. The primary limiting factor is the availability of an external API for both triggering actions (synthesis or measurement), and retrieving results (raw or analyzed data). We encourage researchers and tool vendors to push aggressively towards a world wherein every piece of laboratory equipment has an API, and is thus amenable for AI automation. LLM

technology may in fact be a crucial enabler for such a transition, since their ability to handle arbitrary and heterogeneous APIs (as long as documentation is provided) liberates researchers and manufacturers from having to agree on and follow a single standard for laboratory automation.

Conversely, it must be acknowledged that automation of scientific instruments and laboratory workflows represents a bottleneck for AI-driven science. While AI models and software can be rapidly iterated and improved, hardware system improvements are more capital-intensive and require longer-timescale design and construction efforts. Vendor-provided tools may use proprietary data formats and may not expose software interfaces that provide complete control of the system. These represent significant roadblocks to automation; the community should correspondingly demand commercial solutions that adhere to open data standards. Although mechanizing and automating laboratory work is by no means trivial, we argue that the value of any such effort will increase dramatically in the coming years, as AI agent control systems increase in sophistication.

2.3.2 Open science databases. There is growing appreciation that the data underlying scientific publications should be open and freely available to others. Open data practices increase the realized value of a research effort, as datasets can be used by others in ways not originally envisioned.[96,299] For example, datasets can be used for meta-analysis, to identify broader trends, and as inputs to machine-learning training. The FAIR data principle emphasizes that all datasets should be findable, accessible, interoperable, and reusable.[300] In practice, this means data must be retained and archived, that archives should be open for download and indexing, and that data should be correctly labelled and have corresponding meta-data to contextualize and associate it (with people, groups, publications, and related datasets).

The exocortex is closely tied to open data efforts. To function most effectively, it requires that AI agents be able to identify and operate on vast datasets. Thus, the exocortex is empowered by the greater availability of research data of all types. Obviously, the exocortex also improves as more domain-specific AI modules are trained; this will typically require aggregating openly available domain datasets.

The exocortex concept can also potentially improve data release. One key limiter in data release is researchers being unable to provide sufficiently detailed meta-data, because common tools lack meta-data features and because of the time burden associated with manually adding human annotations to vast datasets. AI agents can help here, as they are better able to handle the ambiguity of sparse meta-data labelling. AI agents may also be able to help automate the collection of meta-data and annotation of datasets when they are first produced, which should increase the richness of meta-data captured. The lesson for the community to learn is that data release is valuable even if the dataset is imperfectly organized and annotated. Future tools will make it possible to organize and extract value even from heterogeneous and unlabelled data.

2.3.3 Software. Software underlies an enormous amount of scientific research, and its importance is further growing as more machine-learning methods are integrated into science. LLMs can play an important role in scientific software, for code generation[5] and code execution by calling APIs[42–50] or interacting with graphic user interfaces (GUIs).[301–304] LLMs could also play a role in user education, since they provide a way for scientists to learn new software systems *via* chatbot assistance or LLM generation of code exemplars.

Greater integration of scientific software tools into AI agent workflows will require these tools to be made readily available. Fortunately, the prevailing trend in scientific software is to release code as open source, and make it available *via* repositories; these make it possible for automated systems such as AI agents to take advantage of them. AI agents may well be able to handle some of the heterogeneity of modern software deployments; that is, they may be able to automatically download code, set up an appropriate containerized environment, generate wrapper code for interacting with that container, and then activate the system. However, it may well be preferable for the community to begin developing and adopting flexible but standardized methods of containerizing scientific software so that it can be more easily shared and launched. For instance, the MLExchange effort is developing a web platform for working with containerized ML models.[305]

We also note a substantial software infrastructure challenge. Running a large number of AI agent instances, and enabling coordination between them, is a substantial technical challenge. Integrating these resources into existing scientific software workflows is also challenging.[6] The expertise of the high-performance computing (HPC) community can be leveraged, as existing know-how with respecting to scaling and deployment of large-scale science software systems should be transferable to AI swarm architectures.

2.3.4 Publications. As with data, publications should ideally be broadly open in order for an exocortex to leverage them. It must be possible for the exocortex to identify relevant documents, retrieve them, and read them. Currently, the vast majority of the scientific literature is not easily available for machine indexing, retrieval, and AI/ML training. Researchers will generally not be able to negotiate the required licenses with publishers in order to obtain access. Luckily, scientific practices have been increasingly moving towards open access, where the publication (or at least a preprint version of it) is freely available. We encourage researchers and publishers to continue pushing in this direction, since it both benefits traditional scientific practices, and helps enable future AI-driven workflows.

In addition to policy questions, there are engineering tasks required in order to make the scientific literature readily available to AI agents. In the short term, researchers may need to manage these activities themselves, building a curated local corpus of documents for their agents to interact with. In the medium term, AI agents will likely be able to access publications through tools or APIs. In the long term, the community would ideally build a common AI database on top of the existing literature. For instance, a community database that kept track of embeddings for every published paper (and sub-sections thereof) would avoid the wasted cost of researchers

recomputing embeddings when their own agent ingests publications. Such a database could also store AI-generated secondary products associated with papers (summaries, classifications, connections to other literature, proposed research directions, *etc.*). Sharing such a database would allow each researcher's exocortex to leverage the work of all other researchers exocortices.

2.3.5 Facilities. Scientific research tools are often organized into coherent facilities that offer multiple related capabilities, or multiple versions of a particular measurement tool (as in the case of electron microscopy centers, synchrotrons, FELs, *etc.*). As more synthesis, processing, and measurement tools become individually automated and collectively organized into facilities, we can begin to imagine the impact that agentic AI will have on them. In particular, agentic AI will enable a transition of scientific facilities away from individual tools that are selected and micro-managed by scientists, and into a discovery ecosystem, wherein users can phrase their high-level scientific goals, and rely upon a swarm of AI agents to correctly select tools, launch experiments, and aggregate results. A possible architecture is depicted in Fig. 5. Each researcher's exocortex, which knows about that researcher's scientific goals and problem-specific science constraints, can negotiate with AI agents operated by the facilities. This allows researchers to conceive of science goals, and leverage agents to convert this into actionable plans. The scientific facilities design and operate AI agents responsible with providing access to a variety of systems, and correctly coordinating between these systems. For instance, a particular research goal might require launching a set of measurement tools and corresponding simulations, and then aggregating the results to compare. This coordination could be executed by a combination of AI agents and traditional software infrastructure. This vision inherently requires ubiquitous and reliable automation of individual systems. It also requires novel developments in research infrastructure, to more efficiently cross-connect between components. We postulate that agentic AI will be an enabling technology for accomplishing this interconnection, as it will bypass the need for every subcomponent to adhere to a single standard for meta-data and communication. As long as each component provides a documented software interface, the layer of AI agents should be able to productively access it. The productivity gain from this architecture could be transformative, as it would allow researchers to conduct experiments of a complexity previously impossible.

3 Perspectives

We have presented an admittedly speculative vision for the future of science, wherein each scientist has a personalized exocortex—a swarm of AI agents working together to automate research and expand researcher cognition. While this vision currently seems far-fetched, it is now within reach owing to recent developments in LLMs; and it becomes increasingly realistic as LLM technology improves. Indeed, the exocortex is envisioned in such a way that it automatically leverages improvements in the technology of AI agents, as more powerful models can be swapped in progressively as they become available. We propose that the science community should work together, and aggressively pursue the creation of systems like this.

We suggest that physical scientists focus on applications of AI agents, and learning how best to connect agents into coherent workflows. In fact, science is an ideal proving ground for agentic AI, since scientists can articulate precise goals, assess rigor of

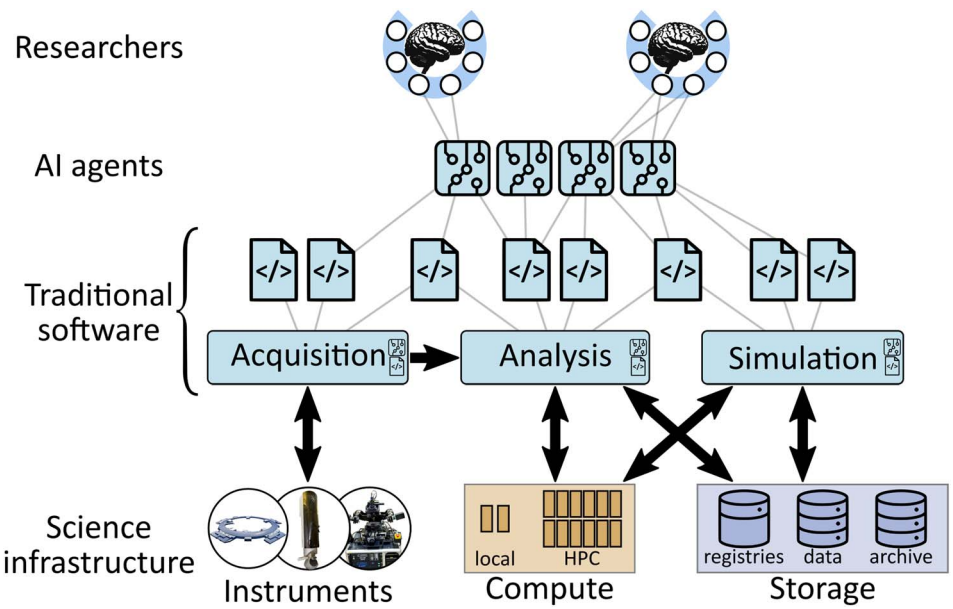

Fig. 5 A possible architecture for AI agents aggregating access to scientific facilities. Each researcher's exocortex could negotiate control by dialoging with a set of AI agents provided by the facilities. That layer of agents would be optimized to launch tasks using traditional software APIs. The underlying resources (measurement instruments, compute resources, databases) would be triggered and queried, with the outputs integrated first by the facility AI agents, and then by the researcher's exocortex.

reasoning, and evaluate success. Thus, AI/ML researchers will hopefully view the physical sciences as an ideal environment in which to research agents and agent swarms.

The proposed multi-agent interactions and workflows highlight several open research questions. It is not known whether complex multi-step AI tasks will be sufficiently robust. The community must measure how AI capabilities scale, as a function of task complexity and inter-agent organizational architecture. The bottlenecks for scientific discovery—especially automated discovery workflows—must be elucidated. We speculate that LLMs will provide high utility for ideation and hypothesis generation, by providing the human with text digests and ranked ideas, and by acting as a conversational partner. However, integration with experimental tools is likely to lag, owing to the time and cost associated with building and testing laboratory automation systems. With respect to the overall exocortex system, we envision the largest roadblocks arising from managing the complexity of inter-communicating agents, and establishing sufficient reliability.

We emphasize that the proposed work is valuable even if the exocortex concept turns out not to be the right framing. The proposed improvements to science infrastructure—making it increasingly robust, automated, software-accessible, and auditable—has value even if AI agents are not successful. The proposed AI agents—streamlining access to publications, data, software, and instruments—are valuable even if their interconnection into an exocortex proves fruitless.

The science exocortex has enormous potential impact. There is growing evidence that generative AI methods exhibit various forms of emergence, including world modeling,[22–25] concept generalization,[25,28–31] and pattern aggregation that is more capable than the inputs.[306] The exocortex architecture would enable and leverage additional layers of emergence. Interactions between AI agents should lead to more reliable, coherent, and capable output than single-shot generation by a lone LLM. And, crucially, interaction between a swarm of AI agents—each responsible for intelligently mediating access to a suite of research capabilities—and a human researcher should lead to the emergence of enhanced human capabilities. By expanding the researcher's intelligence into the exocortex, the researcher can accomplish more, as they are able to intuitively and seamlessly weave myriad physical, computational, and cognititve systems into their intellectual work.

Data availability

As this is a perspective/review article, no primary research results, data, software or code have been included.

Author contributions

K. G. Y. reviewed the literature, developed the concepts, and wrote the manuscript.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This research was carried out by the Center for Functional Nanomaterials, which is a U.S. DOE Office of Science Facility, at Brookhaven National Laboratory under Contract No. DE-SC0012704. We thank Dr Charles T. Black for fruitful discussions.

Notes and references

1 J. Qiu, Q. Wu, G. Ding, Y. Xu and S. Feng, *EURASIP J. Adv. Signal Process.*, 2016, **2016**, 67.

2 Microsoft Research AI4Science and Microsoft Azure Quantum, The Impact of Large Language Models on Scientific Discovery: a Preliminary Study using GPT-4, *arXiv*, 2023, preprint, arXiv:2311.07361, DOI: **10.48550/arXiv.2311.07361**.

3 H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, A. Anandkumar, K. Bergen, C. P. Gomes, S. Ho, P. Kohli, J. Lasenby, J. Leskovec, T.-Y. Liu, A. Manrai, D. Marks, B. Ramsundar, L. Song, J. Sun, J. Tang, P. Veličković, M. Welling, L. Zhang, C. W. Coley, Y. Bengio and M. Zitnik, *Nature*, 2023, **620**, 47–60.

4 T. R. Society, *Science in the Age of AI: How artificial intelligence is changing the nature and method of scientific research*, 2024, **https://royalsociety.org/-/media/policy/projects/science-in-the-age-of-ai/science-in-the-age-of-ai-report.pdf**, accessed: 2024-05-31.

5 K. M. Jablonka, Q. Ai, A. Al-Feghali, S. Badhwar, J. D. Bocarsly, A. M. Bran, S. Bringuier, L. C. Brinson, K. Choudhary, D. Circi, S. Cox, W. A. de Jong, M. L. Evans, N. Gastellu, J. Genzling, M. V. Gil, A. K. Gupta, Z. Hong, A. Imran, S. Kruschwitz, A. Labarre, J. Lála, T. Liu, S. Ma, S. Majumdar, G. W. Merz, N. Moitessier, E. Moubarak, B. Mouriño, B. Pelkie, M. Pieler, M. C. Ramos, B. Ranković, S. G. Rodriques, J. N. Sanders, P. Schwaller, M. Schwarting, J. Shi, B. Smit, B. E. Smith, J. Van Herck, C. Völker, L. Ward, S. Warren, B. Weiser, S. Zhang, X. Zhang, G. A. Zia, A. Scourtas, K. J. Schmidt, I. Foster, A. D. White and B. Blaiszik, *Digital Discovery*, 2023, **2**, 1233–1250.

6 N. C. Hudson, J. G. Pauloski, M. Baughman, A. Kamatar, M. Sakarvadia, L. Ward, R. Chard, A. Bauer, M. Levental, W. Wang, W. Engler, O. Price Skelly, B. Blaiszik, R. Stevens, K. Chard and I. Foster, *Proceedings of the IEEE/ACM 10th International Conference on Big Data Computing, Applications and Technologies*, New York, NY, USA, 2024.

7 T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. 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. 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, *Adv. Neural Inf. Process. Syst.*, 2020, 1877–1901.

8 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. Q. 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. Re, 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, *arXiv*, 2021, preprint, arXiv:2108.07258, DOI: **10.48550/arXiv.2108.07258**.

9 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, Attention Is All You Need, *arXiv*, 2017, preprint, arXiv:1706.03762, DOI: **10.48550/arXiv.1706.03762**.

10 M. Jovanovic and M. Campbell, *Computer*, 2022, **55**, 107–112.

11 R. Gozalo-Brizuela and E. C. Garrido-Merchan, ChatGPT is not all you need. A State of the Art Review of large Generative AI models, *arXiv*, 2023, preprint, arXiv:2301.04655, DOI: **10.48550/arXiv.2301.04655**.

12 A. Ramesh, P. Dhariwal, A. Nichol, C. Chu and M. Chen, Hierarchical Text-Conditional Image Generation with CLIP Latents, *arXiv*, 2022, preprint, arXiv:2204.06125, DOI: **10.48550/arXiv.2204.06125**.

13 R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, High-Resolution Image Synthesis with Latent Diffusion Models, *arXiv*, 2021, preprint, arXiv:2112.10752, DOI: **10.48550/arXiv.2112.10752**.

14 J. Oppenlaender, *Proceedings of the 25th International Academic Mindtrek Conference*, New York, NY, USA, 2022, p. 192–202.

15 A. Radford, K. Narasimhan, T. Salimans and I. Sutskever, *Improving Language Understanding by Generative Pre-Training, Openai Technical Report*, 2018, **https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf**.

16 J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin and X. Hu, Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond, *arXiv*, 2023, preprint, arXiv:2304.13712, DOI: **10.48550/arXiv.2304.13712**.

17 Y. Liu, H. He, T. Han, X. Zhang, M. Liu, J. Tian, Y. Zhang, J. Wang, X. Gao, T. Zhong, Y. Pan, S. Xu, Z. Wu, Z. Liu, X. Zhang, S. Zhang, X. Hu, T. Zhang, N. Qiang, T. Liu and B. Ge, Understanding LLMs: A Comprehensive Overview from Training to Inference, *arXiv*, 2024, preprint, arXiv:2401.02038, DOI: **10.48550/arXiv.2401.02038**.

18 S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain and J. Gao, Large Language Models: A Survey, *arXiv*, 2024, preprint, arXiv:2402.06196, DOI: **10.48550/arXiv.2402.06196**.

19 J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang and Y. Zhou, Deep Learning Scaling is Predictable, Empirically, *arXiv*, 2017, preprint, arXiv:1712.00409, DOI: **10.48550/arXiv.1712.00409**.

20 T. Henighan, J. Kaplan, M. Katz, M. Chen, C. Hesse, J. Jackson, H. Jun, T. B. Brown, P. Dhariwal, S. Gray, C. Hallacy, B. Mann, A. Radford, A. Ramesh, N. Ryder, D. M. Ziegler, J. Schulman, D. Amodei and S. McCandlish, Scaling Laws for Autoregressive Generative Modeling, *arXiv*, 2020, preprint, arXiv:2010.14701, DOI: **10.48550/arXiv.2010.14701**.

21 J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals and L. Sifre, Training Compute-Optimal Large Language Models, *arXiv*, 2022, preprint, arXiv:2203.15556, DOI: **10.48550/arXiv.2203.15556**.

22 K. Li, A. K. Hopkins, D. Bau, F. Viégas, H. Pfister and M. Wattenberg, Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task, *arXiv*, 2023, preprint, arXiv:2210.13382, DOI: **10.48550/arXiv.2210.13382**.

23 E. Akyürek, D. Schuurmans, J. Andreas, T. Ma and D. Zhou, What learning algorithm is in-context learning? Investigations with linear models, *arXiv*, 2023, preprint, arXiv:2211.15661, DOI: **10.48550/arXiv.2211.15661**.

24 M. Kosinski, Evaluating Large Language Models in Theory of Mind Tasks, *arXiv*, 2023, preprint, arXiv:2302.02083, DOI: **10.48550/arXiv.2302.02083**.

25 T. Webb, K. J. Holyoak and H. Lu, *Nat. Human Behav.*, 2023, 1526–1541.

26 W. Gurnee and M. Tegmark, Language Models Represent Space and Time, *arXiv*, 2024, preprint, arXiv:2310.02207, DOI: **10.48550/arXiv.2310.02207**.

27 K. Vafa, J. Y. Chen, J. Kleinberg, S. Mullainathan and A. Rambachan, Evaluating the World Model Implicit in a Generative Model, *arXiv*, 2024, preprint, arXiv:2406.03689, DOI: **10.48550/arXiv.2406.03689**.

28 D. Ganguli, D. Hernandez, L. Lovitt, A. Askell, Y. Bai, A. Chen, T. Conerly, N. Dassarma, D. Drain, N. Elhage, S. E. Showk, S. Fort, Z. Hatfield-Dodds, T. Henighan, S. Johnston, A. Jones, N. Joseph, J. Kernian, S. Kravec, B. Mann, N. Nanda, K. Ndousse, C. Olsson, D. Amodei, T. Brown, J. Kaplan, S. McCandlish, C. Olah, D. Amodei and J. Clark, *ACM Conference on Fairness, Accountability, and Transparency*, 2022.

29 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, *arXiv*, 2022, preprint, arXiv:2206.07682, DOI: **10.48550/arXiv.2206.07682**.

30 N. Nanda, L. Chan, T. Lieberum, J. Smith and J. Steinhardt, Progress measures for grokking via mechanistic interpretability, *arXiv*, 2023, preprint, arXiv:2301.05217, DOI: **10.48550/arXiv.2301.05217**.

31 S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro and Y. Zhang, Sparks of Artificial General Intelligence: Early experiments with GPT-4, *arXiv*, 2023, preprint, arXiv:2303.12712, DOI: **10.48550/arXiv.2303.12712**.

32 D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano and G. Irving, Fine-Tuning Language Models from Human Preferences, *arXiv*, 2020, preprint, arXiv:1909.08593, DOI: **10.48550/arXiv.1909.08593**.

33 L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. 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. Christiano, J. Leike and R. Lowe, Training language models to follow instructions with human feedback, 2022, *arXiv*, preprint, arXiv:2203.02155, DOI: **10.48550/arXiv.2203.02155**.

34 N. Lambert, L. Castricato, L. von Werra and A. Havrilla, *Hugging Face Blog*, 2022.

35 H. Lee, S. Phatale, H. Mansoor, T. Mesnard, J. Ferret, K. Lu, C. Bishop, E. Hall, V. Carbune, A. Rastogi and S. Prakash, RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback, *arXiv*, 2023, preprint, arXiv:2309.00267, DOI: **10.48550/arXiv.2309.00267**.

36 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, *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020.

37 K. G. Yager, *Digital Discovery*, 2023, **2**, 1850–1861.

38 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, *arXiv*, 2024, preprint, arXiv:2312.10997, DOI: **10.48550/arXiv.2312.10997**.

39 H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu and Z. Liu, Evaluation of Retrieval-Augmented Generation: A Survey, *arXiv*, 2024, preprint, arXiv:2405.07437, DOI: **10.48550/arXiv.2405.07437**.

40 A. Karpathy, *@karpathy – Windows, OS X, Linux*, 2023, **https://www.threads.net/@karpathy/post/CzehPtxPEF3**, 11/10/2023.

41 A. Karpathy, *@karpathy – LLM OS*, 2023, **https://www.threads.net/@karpathy/post/CzfH7LQJ7NH**, 11/10/2023.

42 S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan and Y. Cao, ReAct: Synergizing Reasoning and Acting in Language Models, *arXiv*, 2023, preprint, arXiv:2210.03629, DOI: **10.48550/arXiv.2210.03629**.

43 T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda and T. Scialom, Toolformer: Language Models Can Teach Themselves to Use Tools, *arXiv*, 2023, preprint, arXiv:2302.04761, DOI: **10.48550/arXiv.2302.04761**.

44 L. Gao, A. Madaan, S. Zhou, U. Alon, P. Liu, Y. Yang, J. Callan and G. Neubig, PAL: Program-aided Language Models, *arXiv*, 2023, preprint, arXiv:2211.10435, DOI: **10.48550/arXiv.2211.10435**.

45 Y. Liang, C. Wu, T. Song, W. Wu, Y. Xia, Y. Liu, Y. Ou, S. Lu, L. Ji, S. Mao, Y. Wang, L. Shou, M. Gong and N. Duan, TaskMatrix.AI: Completing Tasks by Connecting Foundation Models with Millions of APIs, *arXiv*, 2023, preprint, arXiv:2303.16434, DOI: **10.48550/arXiv.2303.16434**.

46 Y. Shen, K. Song, X. Tan, D. Li, W. Lu and Y. Zhuang, HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face, *arXiv*, 2023, preprint, arXiv:2303.17580, DOI: **10.48550/arXiv.2303.17580**.

47 T. Cai, X. Wang, T. Ma, X. Chen and D. Zhou, Large Language Models as Tool Makers, *arXiv*, 2023, preprint, arXiv:2305.17126, DOI: **10.48550/arXiv.2305.17126**.

48 B. Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Liden, Z. Yu, W. Chen and J. Gao, Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback, *arXiv*, 2023, preprint, arXiv:2302.12813, DOI: **10.48550/arXiv.2302.12813**.

49 B. Xu, Z. Peng, B. Lei, S. Mukherjee, Y. Liu and D. Xu, ReWOO: Decoupling Reasoning from Observations for Efficient Augmented Language Models, *arXiv*, 2023, preprint, arXiv:2305.18323, DOI: **10.48550/arXiv.2305.18323**.

50 C.-Y. Hsieh, S.-A. Chen, C.-L. Li, Y. Fujii, A. Ratner, C.-Y. Lee, R. Krishna and T. Pfister, Tool Documentation Enables Zero-Shot Tool-Usage with Large Language Models, *arXiv*, 2023, preprint, arXiv:2308.00675, DOI: **10.48550/arXiv.2308.00675**.

51 N. Shinn, F. Cassano, B. Labash, A. Gopinath, K. Narasimhan and S. Yao, Reflexion: Language Agents with Verbal Reinforcement Learning, *arXiv*, 2023, preprint, arXiv:2303.11366, DOI: **10.48550/arXiv.2303.11366**.

52 H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever and K. Cobbe, Let's Verify Step by Step, *arXiv*, 2023, preprint, arXiv:2305.20050, DOI: **10.48550/arXiv.2305.20050**.

53 N. McAleese, R. M. Pokorny), J. F. C. Uribe, E. Nitishinskaya, M. Trębacz and J. Leike, *LLM Critics Help Catch LLM Bugs*, 2024, **https://cdn.openai.com/llm-critics-help-catch-llm-bugs-paper.pdf**, accessed: 2024-06-13.

54 W. Xu, A. Banburski-Fahey and N. Jojic, Reprompting: Automated Chain-of-Thought Prompt Inference Through

Gibbs Sampling, *arXiv*, 2023, preprint, arXiv:2305.09993, DOI: 10.48550/arXiv.2305.09993.

55 S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao and K. Narasimhan, Tree of Thoughts: Deliberate Problem Solving with Large Language Models, *arXiv*, 2023, preprint, arXiv:2305.10601, DOI: 10.48550/arXiv.2305.10601.

56 J. Xu, H. Fei, L. Pan, Q. Liu, M.-L. Lee and W. Hsu, Faithful Logical Reasoning via Symbolic Chain-of-Thought, *arXiv*, 2024, preprint, arXiv:2405.18357, DOI: 10.48550/arXiv.2405.18357.

57 G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan and A. Anandkumar, Voyager: An Open-Ended Embodied Agent with Large Language Models, *arXiv*, 2023, preprint, arXiv:2305.16291, DOI: 10.48550/arXiv.2305.16291.

58 G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin and B. Ghanem, CAMEL: Communicative Agents for "Mind" Exploration of Large Scale Language Model Society, *arXiv*, 2023, preprint, arXiv:2303.17760, DOI: 10.48550/arXiv.2303.17760.

59 D. A. Boiko, R. MacKnight, B. Kline and G. Gomes, *Nature*, 2023, **624**, 570–578.

60 R. Yang, J. Chen, Y. Zhang, S. Yuan, A. Chen, K. Richardson, Y. Xiao and D. Yang, SelfGoal: Your Language Agents Already Know How to Achieve High-level Goals, *arXiv*, 2024, preprint, arXiv:2406.04784, DOI: 10.48550/arXiv.2406.04784.

61 T. Bonaci, J. Herron, C. Matlack and H. J. Chizeck, *IEEE Conference on Norbert Wiener in the 21st Century (21CW)*, 2014, pp. 1–8.

62 Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, R. Zheng, X. Fan, X. Wang, L. Xiong, Y. Zhou, W. Wang, C. Jiang, Y. Zou, X. Liu, Z. Yin, S. Dou, R. Weng, W. Cheng, Q. Zhang, W. Qin, Y. Zheng, X. Qiu, X. Huang and T. Gui, The Rise and Potential of Large Language Model Based Agents: A Survey, *arXiv*, 2023, preprint, arXiv:2309.07864, DOI: 10.48550/arXiv.2309.07864.

63 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, *Front. Comput. Sci.*, 2024, **18**, 1–26.

64 M. C. Ramos, C. J. Collison and A. D. White, A Review of Large Language Models and Autonomous Agents in Chemistry, *arXiv*, 2024, preprint, arXiv:2407.01603, DOI: 10.48550/arXiv.2407.01603.

65 H. Jin, L. Huang, H. Cai, J. Yan, B. Li and H. Chen, From LLMs to LLM-based Agents for Software Engineering: A Survey of Current, Challenges and Future, *arXiv*, 2024, preprint, arXiv:2408.02479, DOI: 10.48550/arXiv.2408.02479.

66 S. Kapoor, B. Stroebl, Z. S. Siegel, N. Nadgir and A. Narayanan, AI Agents That Matter, *arXiv*, 2024, preprint, arXiv:2407.01502, DOI: 10.48550/arXiv.2407.01502.

67 W. Zhong, L. Guo, Q. Gao, H. Ye and Y. Wang, MemoryBank: Enhancing Large Language Models with Long-Term Memory, *arXiv*, 2023, preprint, arXiv:2305.10250, DOI: 10.48550/arXiv.2305.10250.

68 W. Wang, L. Dong, H. Cheng, X. Liu, X. Yan, J. Gao and F. Wei, Augmenting Language Models with Long-Term Memory, *arXiv*, 2023, preprint, arXiv:2306.07174, DOI: 10.48550/arXiv.2306.07174.

69 P. Das, S. Chaudhury, E. Nelson, I. Melnyk, S. Swaminathan, S. Dai, A. Lozano, G. Kollias, V. Chenthamarakshan, N. Jiří, S. Dan and P.-Y. Chen, Larimar: Large Language Models with Episodic Memory Control, 2024, *arXiv*, preprint, arXiv:2403.11901, DOI: 10.48550/arXiv.2403.11901.

70 J. Li, S. Consul, E. Zhou, J. Wong, N. Farooqui, N. Manohar, Z. N. Wei, T. Wu, B. Echols, S. Zhou and G. Diamos, *Banishing LLM Hallucinations Requires Rethinking Generalization, github*, 2024, https://github.com/lamini-ai/Lamini-Memory-Tuning/blob/main/research-paper.pdf, accessed: 2024-06-13.

71 E. Zelikman, G. Harik, Y. Shao, V. Jayasiri, N. Haber and N. D. Goodman, Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking, *arXiv*, 2024, preprint, arXiv:2403.09629, DOI: 10.48550/arXiv.2403.09629.

72 W. Bounsi, B. Ibarz, A. Dudzik, J. B. Hamrick, L. Markeeva, A. Vitvitskyi, R. Pascanu and P. Veličković, Transformers meet Neural Algorithmic Reasoners, *arXiv*, 2024, preprint, arXiv:2406.09308, DOI: 10.48550/arXiv.2406.09308.

73 L. Luo, Y. Liu, R. Liu, S. Phatale, H. Lara, Y. Li, L. Shu, Y. Zhu, L. Meng, J. Sun and A. Rastogi, Improve Mathematical Reasoning in Language Models by Automated Process Supervision, *arXiv*, 2024, preprint, arXiv:2406.06592, DOI: 10.48550/arXiv.2406.06592.

74 G. Chen, M. Liao, C. Li and K. Fan, AlphaMath Almost Zero: process Supervision without process, *arXiv*, 2024, preprint, arXiv:2405.03553, DOI: 10.48550/arXiv.2405.03553.

75 D. Zhang, S. Zhoubian, Y. Yue, Y. Dong and J. Tang, ReST-MCTS*: LLM Self-Training via Process Reward Guided Tree Search, *arXiv*, 2024, preprint, arXiv:2406.03816, DOI: 10.48550/arXiv.2406.03816.

76 D. Zhang, J. Li, X. Huang, D. Zhou, Y. Li and W. Ouyang, Accessing GPT-4 level Mathematical Olympiad Solutions via Monte Carlo Tree Self-refine with LLaMa-3 8B, *arXiv*, 2024, preprint, arXiv:2406.07394, DOI: 10.48550/arXiv.2406.07394.

77 J. Y. Koh, S. McAleer, D. Fried and R. Salakhutdinov, *arXiv*, 2024, preprint, arXiv:2407.01476, DOI: 10.48550/arXiv.2407.01476.

78 C. Li, W. Wang, J. Hu, Y. Wei, N. Zheng, H. Hu, Z. Zhang and H. Peng, Common 7B Language Models Already Possess Strong Math Capabilities, *arXiv*, 2024, preprint, arXiv:2403.04706, DOI: 10.48550/arXiv.2403.04706.

79 B.-W. Zhang, Y. Yan, L. Li and G. Liu, InfinityMATH: A Scalable Instruction Tuning Dataset in Programmatic Mathematical Reasoning, *arXiv*, 2024, preprint, arXiv:2408.07089, DOI: 10.48550/arXiv.2408.07089.

80 S. Golkar, M. Pettee, M. Eickenberg, A. Bietti, M. Cranmer, G. Krawezik, F. Lanusse, M. McCabe, R. Ohana, L. Parker,

B. R.-S. Blancard, T. Tesileanu, K. Cho and S. Ho, xVal: A Continuous Number Encoding for Large Language Models, *arXiv*, 2023, preprint, arXiv:2310.02989, DOI: 10.48550/arXiv.2310.02989.

81 S. McLeish, A. Bansal, A. Stein, N. Jain, J. Kirchenbauer, B. R. Bartoldson, B. Kailkhura, A. Bhatele, J. Geiping, A. Schwarzschild and T. Goldstein, Transformers Can Do Arithmetic with the Right Embeddings, *arXiv*, 2024, preprint, arXiv:2405.17399, DOI: 10.48550/arXiv.2405.17399.

82 T. H. Trinh, Y. Wu, Q. V. Le, H. He and T. Luong, *Nature*, 2024, **625**, 476–482.

83 A. Vashishtha, A. Kumar, A. G. Reddy, V. N. Balasubramanian and A. Sharma, Teaching Transformers Causal Reasoning through Axiomatic Training, *arXiv*, 2024, preprint, arXiv:2407.07612, DOI: 10.48550/arXiv.2407.07612.

84 P. Emami, Z. Li, S. Sinha and T. Nguyen, SysCaps: Language Interfaces for Simulation Surrogates of Complex Systems, *arXiv*, 2024, preprint, arXiv:2405.19653, DOI: 10.48550/arXiv.2405.19653.

85 S. Kantamneni, Z. Liu and M. Tegmark, OptPDE: Discovering Novel Integrable Systems via AI-Human Collaboration, *arXiv*, 2024, preprint, arXiv:2405.04484, DOI: 10.48550/arXiv.2405.04484.

86 V. Kumar, L. Gleyzer, A. Kahana, K. Shukla and G. E. Karniadakis, *J. Mach. Learn. Model. Comput.*, 2023, **4**, 41–72.

87 S. Jia, C. Zhang and V. Fung, LLMatDesign: Autonomous Materials Discovery with Large Language Models, *arXiv*, 2024, preprint, arXiv:2406.13163, DOI: 10.48550/arXiv.2406.13163.

88 P. M. Maffettone, L. Banko, P. Cui, Y. Lysogorskiy, M. A. Little, D. Olds, A. Ludwig and A. I. Cooper, *Nat. Comput. Sci.*, 2021, **1**, 290–297.

89 G. A. Tritsaris, S. Carr and G. R. Schleder, *Applied Physics Reviews*, 2021, **8**, 031401.

90 A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White and P. Schwaller, ChemCrow: Augmenting large-language models with chemistry tools, *arXiv*, 2023, preprint, arXiv:2304.05376, DOI: 10.48550/arXiv.2304.05376.

91 M. Xu, X. Yuan, S. Miret and J. Tang, ProtST: Multi-Modality Learning of Protein Sequences and Biomedical Texts, *arXiv*, 2023, preprint, arXiv:2301.12040, DOI: 10.48550/arXiv.2301.12040.

92 S. Liu, Y. Li, Z. Li, A. Gitter, Y. Zhu, J. Lu, Z. Xu, W. Nie, A. Ramanathan, C. Xiao, J. Tang, H. Guo and A. Anandkumar, A Text-guided Protein Design Framework, *arXiv*, 2023, preprint, arXiv:2302.04611, DOI: 10.48550/arXiv.2302.04611.

93 M. Y. Lu, B. Chen, D. F. K. Williamson, R. J. Chen, M. Zhao, A. K. Chow, K. Ikemura, A. Kim, D. Pouli, A. Patel, A. Soliman, C. Chen, T. Ding, J. J. Wang, G. Gerber, I. Liang, L. P. Le, A. V. Parwani, L. L. Weishaupt and F. Mahmood, *Nature*, 2024, DOI: 10.1038/s41586-024-07618-3.

94 L. A. Royer, *Nat. Methods*, 2024, **21**(8), 1371–1373.

95 K. G. Yager, *Online Resource for Big Data and Extreme-Scale Computing Workshop*, 2018.

96 K. G. Yager, *Methods and Applications of Autonomous Experimentation*, Chapman and Hall/CRC, 2023, 1st edn, ch. 1, p. 21.

97 L. Hung, J. A. Yager, D. Monteverde, D. Baiocchi, H.-K. Kwon, S. Sun and S. Suram, *Digital Discovery*, 2024, **3**(7), 1273–1279.

98 R. D. King, K. E. Whelan, F. M. Jones, P. G. K. Reiser, C. H. Bryant, S. H. Muggleton, D. B. Kell and S. G. Oliver, *Nature*, 2004, **427**, 247–252.

99 A. G. Kusne, T. Gao, A. Mehta, L. Ke, M. C. Nguyen, K.-M. Ho, V. Antropov, C.-Z. Wang, M. J. Kramer, C. Long and I. Takeuchi, *Sci. Rep.*, 2014, **4**, 6367.

100 P. Nikolaev, D. Hooper, N. Perea-López, M. Terrones and B. Maruyama, *ACS Nano*, 2014, **8**, 10214–10222.

101 D. Xue, P. V. Balachandran, R. Yuan, T. Hu, X. Qian, E. R. Dougherty and T. Lookman, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**, 13301–13306.

102 F. Ren, L. Ward, T. Williams, K. J. Laws, C. Wolverton, J. Hattrick-Simpers and A. Mehta, *Sci. Adv.*, 2018, **4**, eaaq1566.

103 H. S. Stein and J. M. Gregoire, *Chem. Sci.*, 2019, **10**, 9640–9649.

104 M. M. Noack, K. G. Yager, M. Fukuto, G. S. Doerk, R. Li and J. A. Sethian, *Sci. Rep.*, 2019, **9**, 11809.

105 M. M. Noack, G. S. Doerk, R. Li, M. Fukuto and K. G. Yager, *Sci. Rep.*, 2020, **10**, 1325.

106 M. M. Noack, G. S. Doerk, R. Li, J. K. Streit, R. A. Vaia, K. G. Yager and M. Fukuto, *Sci. Rep.*, 2020, **10**, 17663.

107 M. M. Noack, P. H. Zwart, D. M. Ushizima, M. Fukuto, K. G. Yager, K. C. Elbert, C. B. Murray, A. Stein, G. S. Doerk, E. H. R. Tsai, R. Li, G. Freychet, M. Zhernenkov, H.-Y. N. Holman, S. Lee, L. Chen, E. Rotenberg, T. Weber, Y. L. Goc, M. Boehm, P. Steffens, P. Mutti and J. A. Sethian, *Nat. Rev. Phys.*, 2021, **3**, 685–697.

108 S. V. Kalinin, M. Ziatdinov, J. Hinkle, S. Jesse, A. Ghosh, K. P. Kelley, A. R. Lupini, B. G. Sumpter and R. K. Vasudevan, *ACS Nano*, 2021, **15**, 12604–12627.

109 E. Stach, B. DeCost, A. G. Kusne, J. Hattrick-Simpers, K. A. Brown, K. G. Reyes, J. Schrier, S. Billinge, T. Buonassisi, I. Foster, C. P. Gomes, J. M. Gregoire, A. Mehta, J. Montoya, E. Olivetti, C. Park, E. Rotenberg, S. K. Saikin, S. Smullin, V. Stanev and B. Maruyama, *Matter*, 2021, **4**, 2702–2726.

110 I.-J. Chen, M. Aapro, A. Kipnis, A. Ilin, P. Liljeroth and A. S. Foster, *Nat. Commun.*, 2022, **13**, 7499.

111 C. Zhao, C.-C. Chung, S. Jiang, M. M. Noack, J.-H. Chen, K. Manandhar, J. Lynch, H. Zhong, W. Zhu, P. Maffettone, D. Olds, M. Fukuto, I. Takeuchi, S. Ghose, T. Caswell, K. G. Yager and Y.-c. K. Chen-Wiegart, *Commun. Mater.*, 2022, **3**, 86.

112 G. S. Doerk, A. Stein, S. Bae, M. M. Noack, M. Fukuto and K. G. Yager, *Sci. Adv.*, 2023, **9**, eadd3687.

113 S. Bae, M. M. Noack and K. G. Yager, *Nanoscale*, 2023, **15**, 6901–6912.

114 K. G. Yager, P. W. Majewski, M. M. Noack and M. Fukuto, *Nanotechnology*, 2023, **34**, 322001.
115 A. A. Volk, R. W. Epps, D. T. Yonemoto, B. S. Masters, F. N. Castellano, K. G. Reyes and M. Abolhasani, *Nat. Commun.*, 2023, **14**, 1403.
116 N. J. Szymanski, B. Rendy, Y. Fei, R. E. Kumar, T. He, D. Milsted, M. J. McDermott, M. Gallant, E. D. Cubuk, A. Merchant, H. Kim, A. Jain, C. J. Bartel, K. Persson, Y. Zeng and G. Ceder, *Nature*, 2023, **624**, 86–91.
117 F. J. Alexander, J. Ang, J. A. Bilbrey, J. Balewski, T. Casey, R. Chard, J. Choi, S. Choudhury, B. Debusschere, A. M. DeGennaro, N. Dryden, J. A. Ellis, I. Foster, C. G. Cardona, S. Ghosh, P. Harrington, Y. Huang, S. Jha, T. Johnston, A. Kagawa, R. Kannan, N. Kumar, Z. Liu, N. Maruyama, S. Matsuoka, E. McCarthy, J. Mohd-Yusof, P. Nugent, Y. Oyama, T. Proffen, D. Pugmire, S. Rajamanickam, V. Ramakrishniah, M. Schram, S. K. Seal, G. Sivaraman, C. Sweeney, L. Tan, R. Thakur, B. V. Essen, L. Ward, P. Welch, M. Wolf, S. S. Xantheas, K. G. Yager, S. Yoo and B.-J. Yoon, *Int. J. High Perform. Comput. Appl.*, 2021, **35**, 598–616.
118 T. Rainforth, A. Foster, D. R. Ivanova and F. B. Smith, Modern Bayesian Experimental Design, *arXiv*, 2023, preprint, arXiv:2302.14545, DOI: **10.1214/23-STS915**.
119 M. M. Noack, *Methods and Applications of Autonomous Experimentation*, Chapman and Hall/CRC, 2023, edn. 1st, ch. 4, p. 16.
120 P. M. Maffettone, D. B. Allan, S. I. Campbell, M. R. Carbone, T. A. Caswell, B. L. DeCost, D. Gavrilov, M. D. Hanwell, H. Joress, J. Lynch, B. Ravel, S. B. Wilkins, J. Wlodek and D. Olds, Self-driving Multimodal Studies at User Facilities, *arXiv*, 2023, preprint, arXiv:2301.09177, DOI: **10.48550/arXiv.2301.09177**.
121 P. Zahl, T. Wagner, R. Möller and A. Klust, *J. Vac. Sci. Technol. B*, 2010, **28**, C4E39–C4E47.
122 Y. Liu, K. P. Kelley, R. K. Vasudevan, H. Funakubo, M. A. Ziatdinov and S. V. Kalinin, *Nat. Mach. Intell.*, 2022, **4**, 341–350.
123 J. Hill, S. Campbell, G. Carini, Y.-C. K. Chen-Wiegart, Y. Chu, A. Fluerasu, M. Fukuto, M. Idir, J. Jakoncic, I. Jarrige, P. Siddons, T. Tanabe and K. G. Yager, *J. Phys.: Condens. Matter*, 2020, **32**, 374008.
124 C. Bostedt, S. Boutet, D. M. Fritz, Z. Huang, H. J. Lee, H. T. Lemke, A. Robert, W. F. Schlotter, J. J. Turner and G. J. Williams, *Rev. Mod. Phys.*, 2016, **88**, 015007.
125 C. Bostedt, H. N. Chapman, J. T. Costello, J. R. Crespo López-Urrutia, S. Düsterer, S. W. Epp, J. Feldhaus, A. Föhlisch, M. Meyer, T. Möller, R. Moshammer, M. Richter, K. Sokolowski-Tinten, A. Sorokin, K. Tiedtke, J. Ullrich and W. Wurth, *Nucl. Instrum. Methods Phys. Res. Sect. A Accel. Spectrom. Detect. Assoc. Equip.*, 2009, **601**, 108–122.
126 E. Allaria, R. Appio, L. Badano, W. A. Barletta, S. Bassanese, S. G. Biedron, A. Borga, E. Busetto, D. Castronovo, P. Cinquegrana, S. Cleva, D. Cocco, M. Cornacchia, P. Craievich, I. Cudin, G. D'Auria, M. Dal Forno, M. B. Danailov, R. De Monte, G. De Ninno, P. Delgiusto, A. Demidovich, S. Di Mitri, B. Diviacco, A. Fabris, R. Fabris, W. Fawley, M. Ferianis, E. Ferrari, S. Ferry, L. Froehlich, P. Furlan, G. Gaio, F. Gelmetti, L. Giannessi, M. Giannini, R. Gobessi, R. Ivanov, E. Karantzoulis, M. Lonza, A. Lutman, B. Mahieu, M. Milloch, S. V. Milton, M. Musardo, I. Nikolov, S. Noe, F. Parmigiani, G. Penco, M. Petronio, L. Pivetta, M. Predonzani, F. Rossi, L. Rumiz, A. Salom, C. Scafuri, C. Serpico, P. Sigalotti, S. Spampinati, C. Spezzani, M. Svandrlik, C. Svetina, S. Tazzari, M. Trovo, R. Umer, A. Vascotto, M. Veronese, R. Visintini, M. Zaccaria, D. Zangrando and M. Zangrando, *Nat. Photonics*, 2012, **6**, 699–704.
127 E. M. Chan, C. Xu, A. W. Mao, G. Han, J. S. Owen, B. E. Cohen and D. J. Milliron, *Nano Lett.*, 2010, **10**, 1874–1885.
128 B. Li, S. S. Kaye, C. Riley, D. Greenberg, D. Galang and M. S. Bailey, *ACS Comb. Sci.*, 2012, **14**, 352–358.
129 Q. Yan, J. Yu, S. K. Suram, L. Zhou, A. Shinde, P. F. Newhouse, W. Chen, G. Li, K. A. Persson, J. M. Gregoire and J. B. Neaton, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, 3040–3043.
130 J. M. Granda, L. Donina, V. Dragone, D.-L. Long and L. Cronin, *Nature*, 2018, **559**, 377–381.
131 R. Vescovi, T. Ginsburg, K. Hippe, D. Ozgulbas, C. Stone, A. Stroka, R. Butler, B. Blaiszik, T. Brettin, K. Chard, M. Hereld, A. Ramanathan, R. Stevens, A. Vriza, J. Xu, Q. Zhang and I. Foster, *Digital Discovery*, 2023, **2**, 1980–1998.
132 R. Shimizu, S. Kobayashi, Y. Watanabe, Y. Ando and T. Hitosugi, *APL Mater.*, 2020, **8**, 111110.
133 M. Abolhasani and E. Kumacheva, *Nat., Synth.*, 2023, **2**(6), 483–492.
134 A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu and B. Zitkovich, *arXiv*, 2023, preprint arXiv:2307.15818, DOI: **10.48550/arXiv.2307.15818**.
135 C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel and S. Song, *Proceedings of Robotics: Science and Systems (RSS)*, 2023.
136 A. Avetisyan, C. Xie, H. Howard-Jenkins, T.-Y. Yang, S. Aroudj, S. Patra, F. Zhang, D. Frost, L. Holland, C. Orme, J. Engel, E. Miller, R. Newcombe and V. Balntas, SceneScript: Reconstructing Scenes With An Autoregressive Structured Language Model, *arXiv*, 2024, preprint, arXiv:2403.13064, DOI: **10.48550/arXiv.2403.13064**.

137 I. Radosavovic, B. Zhang, B. Shi, J. Rajasegaran, S. Kamat, T. Darrell, K. Sreenath and J. Malik, Humanoid Locomotion as Next Token Prediction, *arXiv*, 2024, preprint, arXiv:2402.19469, DOI: **10.48550/arXiv.2402.19469**.

138 M. Ahn, D. Dwibedi, C. Finn, M. G. Arenas, K. Gopalakrishnan, K. Hausman, B. Ichter, A. Irpan, N. Joshi, R. Julian, S. Kirmani, I. Leal, E. Lee, S. Levine, Y. Lu, I. Leal, S. Maddineni, K. Rao, D. Sadigh, P. Sanketi, P. Sermanet, Q. Vuong, S. Welker, F. Xia, T. Xiao, P. Xu, S. Xu and Z. Xu, AutoRT: Embodied Foundation Models for Large Scale Orchestration of Robotic Agents, *arXiv*, 2024, preprint, arXiv:2401.12963, DOI: **10.48550/arXiv.2401.12963**.

139 D. Aldarondo, J. Merel, J. D. Marshall, L. Hasenclever, U. Klibaite, A. Gellis, Y. Tassa, G. Wayne, M. Botvinick and B. P. Ölveczky, *Nature*, 2024, **632**(8025), 594–602.

140 Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein and C. Finn, *arXiv*, 2024, preprint, arXiv:2406.10454, DOI: **10.48550/arXiv.2406.10454**.

141 M. H. Prince, H. Chan, A. Vriza, T. Zhou, V. K. Sastry, M. T. Dearing, R. J. Harder, R. K. Vasudevan and M. J. Cherukara, Opportunities for Retrieval and Tool Augmented Large Language Models in Scientific Facilities, *arXiv*, 2023, preprint, arXiv:2312.01291, DOI: **10.48550/arXiv.2312.01291**.

142 D. Potemkin, C. Soto, R. Li, K. Yager and E. Tsai, Virtual Scientific Companion for Synchrotron Beamlines: A Prototype, *arXiv*, 2023, preprint, arXiv:2312.17180, DOI: **10.48550/arXiv.2312.17180**.

143 Y. Liu, M. Checa and R. K. Vasudevan, Synergizing Human Expertise and AI Efficiency with Language Model for Microscopy Operation and Automated Experiment Design, *arXiv*, 2024, preprint, arXiv:2401.13803, DOI: **10.48550/arXiv.2401.13803**.

144 A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, *arXiv*, 2021, preprint, arXiv:2103.00020, DOI: **10.48550/arXiv.2103.00020**.

145 J. Lu, D. Batra, D. Parikh and S. Lee, ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks, *arXiv*, 2019, preprint, arXiv:1908.02265, DOI: **10.48550/arXiv.1908.02265**.

146 Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu and L. Wang, The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision), *arXiv*, 2023, preprint, arXiv:2309.17421, DOI: **10.48550/arXiv.2309.17421**.

147 K. Carolan, L. Fennelly and A. F. Smeaton, A Review of Multi-Modal Large Language and Vision Models, *arXiv*, 2024, preprint, arXiv:2404.01322, DOI: **10.48550/arXiv.2404.01322**.

148 W. Gao, Z. Deng, Z. Niu, F. Rong, C. Chen, Z. Gong, W. Zhang, D. Xiao, F. Li, Z. Cao, Z. Ma, W. Wei and L. Ma, OphGLM: Training an Ophthalmology Large Language-and-Vision Assistant based on Instructions and Dialogue, *arXiv*, 2023, preprint, arXiv:2306.12174, DOI: **10.48550/arXiv.2306.12174**.

149 C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon and J. Gao, *Adv. Neural Inf. Process. Syst.*, 2023, 28541–28564.

150 Y. Wang, W. Zhang, S. Lin, M. S. Farruggio and A. Wang, *bioRxiv*, 2024, preprint, 2024.04.11.588958, DOI: **10.1101/2024.04.11.588958**.

151 R. Chen, T. Zhao, A. Jaiswal, N. Shah and Z. Wang, LLaGA: Large Language and Graph Assistant, *arXiv*, 2024, preprint, arXiv:2402.08170, DOI: **10.48550/arXiv.2402.08170**.

152 Z. Song, Y. Li, M. Fang, Z. Chen, Z. Shi, Y. Huang and L. Chen, MMAC-Copilot: Multi-modal Agent Collaboration Operating System Copilot, *arXiv*, 2024, preprint, arXiv:2404.18074v2, DOI: **10.48550/arXiv.2404.18074**.

153 P. W. Majewski and K. G. Yager, *J. Phys.: Condens. Matter*, 2016, **28**, 403002.

154 D. Mizrahi, R. Bachmann, O. F. Kar, T. Yeo, M. Gao, A. Dehghan and A. Zamir, *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

155 R. Bachmann, O. F. Kar, D. Mizrahi, A. Garjani, M. Gao, D. Griffiths, J. Hu, A. Dehghan and A. Zamir, 4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities, *arXiv*, 2024, preprint, arXiv:2406.09406, DOI: **10.48550/arXiv.2406.09406**.

156 A. I. Polymathic, *Advancing Science through Multi-Disciplinary AI*, 2024, **https://polymathic-ai.org/**, accessed: 2024-06-10.

157 M. Cranmer, *The Next Great Scientific Theory is Hiding Inside a Neural Network*, 2024, **https://www.simonsfoundation.org/event/the-next-great-scientific-theory-is-hiding-inside-a-neural-network/**, **https://www.youtube.com/watch?v=fk2r8y5TfNY**, Simons Foundation Presidential Lecture, YouTube.

158 F. Lanusse, L. Parker, S. Golkar, M. Cranmer, A. Bietti, M. Eickenberg, G. Krawezik, M. McCabe, R. Ohana, M. Pettee, B. R.-S. Blancard, T. Tesileanu, K. Cho and S. Ho, AstroCLIP: Cross-Modal Pre-Training for Astronomical Foundation Models, *arXiv*, 2023, preprint, arXiv:2310.03024, DOI: **10.1093/mnras/stae1450**.

159 M. McCabe, B. R.-S. Blancard, L. Parker, R. Ohana, M. Cranmer, A. Bietti, M. Eickenberg, S. Golkar, G. Krawezik, F. Lanusse, M. Pettee, T. Tesileanu, K. Cho and S. Ho, *NeurIPS 2023 AI for Science Workshop*, 2023.

160 J. Treutlein, D. Choi, J. Betley, C. Anil, S. Marks, R. B. Grosse and O. Evans, Connecting the Dots: LLMs can Infer and Verbalize Latent Structure from Disparate Training Data, *arXiv*, 2024, preprint, arXiv:2406.14546, DOI: **10.48550/arXiv.2406.14546**.

161 M. M. Noack and J. A. Sethian, *Commun. Appl. Math. Comput. Sci.*, 2021, **17**, 131–156.

162 M. M. Noack, D. Perryman, H. Krishnan and P. H. Zwart, *3rd Annual Workshop on Extreme-scale Experiment-in-the-Loop Computing*, XLOOP, 2021, pp. 24–29.

163 M. M. Noack, H. Krishnan, M. D. Risser and K. G. Reyes, *Sci. Rep.*, 2023, **13**, 3155.

164 B. Poole, A. Jain, J. T. Barron and B. Mildenhall, DreamFusion: Text-to-3D using 2D Diffusion, *arXiv*, 2022, preprint, arXiv:2209.14988, DOI: **10.48550/arXiv.2209.14988**.

165 Z. Wang, C. Lu, Y. Wang, F. Bao, C. Li, H. Su and J. Zhu, ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation, *arXiv*, 2023, preprint, arXiv:2305.16213, DOI: **10.48550/arXiv.2305.16213**.

166 C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu and T.-Y. Lin, Magic3D: High-Resolution Text-to-3D Content Creation, *arXiv*, 2023, preprint, arXiv:2211.10440, DOI: **10.48550/arXiv.2211.10440**.

167 G. Metzer, E. Richardson, O. Patashnik, R. Giryes and D. Cohen-Or, Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures, *arXiv*, 2022, preprint, arXiv:2211.07600, DOI: **10.48550/arXiv.2211.07600**.

168 R. Chen, Y. Chen, N. Jiao and K. Jia, Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation, *arXiv*, 2023, preprint, arXiv:2303.13873, DOI: **10.48550/arXiv.2303.13873**.

169 C. Tsalicoglou, F. Manhardt, A. Tonioni, M. Niemeyer and F. Tombari, TextMesh: Generation of Realistic 3D Meshes From Text Prompts, *arXiv*, 2023, preprint, arXiv:2304.12439, DOI: **10.48550/arXiv.2304.12439**.

170 R. Liu, R. Wu, B. V. Hoorick, P. Tokmakov, S. Zakharov and C. Vondrick, Zero-1-to-3: Zero-shot One Image to 3D Object, *arXiv*, 2023, preprint, arXiv:2303.11328, DOI: **10.48550/arXiv.2303.11328**.

171 G. Qian, J. Mai, A. Hamdi, J. Ren, A. Siarohin, B. Li, H.-Y. Lee, I. Skorokhodov, P. Wonka, S. Tulyakov and B. Ghanem, Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors, 2023, *arXiv*, preprint, arXiv:2306.17843, DOI: **10.48550/arXiv.2306.17843**.

172 A. Haque, M. Tancik, A. A. Efros, A. Holynski and A. Kanazawa, Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions, *arXiv*, 2023, preprint, arXiv:2303.12789, DOI: **10.48550/arXiv.2303.12789**.

173 R. Gao, A. Holynski, P. Henzler, A. Brussee, R. Martin-Brualla, P. Srinivasan, J. T. Barron and B. Poole, CAT3D: Create Anything in 3D with Multi-View Diffusion Models, *arXiv*, 2024, preprint, arXiv:2405.10314, DOI: **10.48550/arXiv.2405.10314**.

174 U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta and Y. Taigman, Make-A-Video: Text-to-Video Generation without Text-Video Data, *arXiv*, 2022, preprint, arXiv:2209.14792, DOI: **10.48550/arXiv.2209.14792**.

175 J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet and T. Salimans, Imagen Video: High Definition Video Generation with Diffusion Models, *arXiv*, 2022, preprint, arXiv:2210.02303, DOI: **10.48550/arXiv.2210.02303**.

176 A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler and K. Kreis, Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models, *arXiv*, 2023, preprint, arXiv:2304.08818, DOI: **10.48550/arXiv.2304.08818**.

177 A. Gupta, L. Yu, K. Sohn, X. Gu, M. Hahn, L. Fei-Fei, I. Essa, L. Jiang and J. Lezama, Photorealistic Video Generation with Diffusion Models, *arXiv*, 2023, preprint, arXiv:2312.06662, DOI: **10.48550/arXiv.2312.06662**.

178 D. Kondratyuk, L. Yu, X. Gu, J. Lezama, J. Huang, G. Schindler, R. Hornung, V. Birodkar, J. Yan, M.-C. Chiu, K. Somandepalli, H. Akbari, Y. Alon, Y. Cheng, J. Dillon, A. Gupta, M. Hahn, A. Hauth, D. Hendon, A. Martinez, D. Minnen, M. Sirotenko, K. Sohn, X. Yang, H. Adam, M.-H. Yang, I. Essa, H. Wang, D. A. Ross, B. Seybold and L. Jiang, VideoPoet: A Large Language Model for Zero-Shot Video Generation, *arXiv*, 2024, preprint, arXiv:2312.14125, DOI: **10.48550/arXiv.2312.14125**.

179 T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang and A. Ramesh, *Video generation models as world simulators*, 2024, **https://openai.com/research/video-generation-models-as-world-simulators**, accessed: 2024-02-15.

180 B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi and R. Ng, *Computer Vision – ECCV*, Cham, 2020, pp. 405–421.

181 B. Kerbl, G. Kopanas, T. Leimkühler and G. Drettakis, 3D Gaussian Splatting for Real-Time Radiance Field Rendering, *arXiv*, 2023, preprint, arXiv:2308.04079, DOI: **10.48550/arXiv.2308.04079**.

182 G. Wu, T. Yi, J. Fang, L. Xie, X. Zhang, W. Wei, W. Liu, Q. Tian and X. Wang, 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering, *arXiv*, 2023, preprint, arXiv:2310.08528, DOI: **10.48550/arXiv.2310.08528**.

183 Z. Li, Z. Chen, Z. Li and Y. Xu, Spacetime Gaussian Feature Splatting for Real-Time Dynamic View Synthesis, *arXiv*, 2024, preprint, arXiv:2312.16812, DOI: **10.48550/arXiv.2312.16812**.

184 J. Ren, K. Xie, A. Mirzaei, H. Liang, X. Zeng, K. Kreis, Z. Liu, A. Torralba, S. Fidler, S. W. Kim and H. Ling, L4GM: Large 4D Gaussian Reconstruction Model, *arXiv*, 2024, preprint, arXiv:2406.10324, DOI: **10.48550/arXiv.2406.10324**.

185 R. Shao, J. Sun, C. Peng, Z. Zheng, B. Zhou, H. Zhang and Y. Liu, Control4D: Efficient 4D Portrait Editing with Text, *arXiv*, 2023, preprint, arXiv:2305.20082, DOI: **10.48550/arXiv.2305.20082**.

186 S. Peng, Y. Zhang and K. Li, PAPR in Motion: Seamless Point-level 3D Scene Interpolation, *arXiv*, 2024, preprint, arXiv:2406.05533, DOI: **10.48550/arXiv.2406.05533**.

187 H. Yu, C. Wang, P. Zhuang, W. Menapace, A. Siarohin, J. Cao, L. A. Jeni, S. Tulyakov and H.-Y. Lee, 4Real: Towards Photorealistic 4D Scene Generation via Video Diffusion Models, *arXiv*, 2024, preprint, arXiv:2406.07472, DOI: **10.48550/arXiv.2406.07472**.

188 Y. Wang, X. Wang, Z. Chen, Z. Wang, F. Sun and J. Zhu, Vidu4D: Single Generated Video to High-Fidelity 4D Reconstruction with Dynamic Gaussian Surfels, *arXiv*,

2024, preprint, arXiv:2405.16822, DOI: 10.48550/arXiv.2405.16822.
189 H. Pang, H. Zhu, A. Kortylewski, C. Theobalt and M. Habermann, ASH: Animatable Gaussian Splats for Efficient and Photoreal Human Rendering, *arXiv*, 2024, preprint, arXiv:2312.05941, DOI: 10.48550/arXiv.2312.05941.
190 D. Duckworth, P. Hedman, C. Reiser, P. Zhizhin, J.-F. Thibert, M. Lučić, R. Szeliski and J. T. Barron, SMERF: Streamable Memory Efficient Radiance Fields for Real-Time Large-Scene Exploration, *arXiv*, 2024, preprint, arXiv:2312.07541, DOI: 10.48550/arXiv.2312.07541.
191 Z. Peng, T. Shao, L. Yong, J. Zhou, Y. Yang, J. Wang and K. Zhou, X-SLAM: Scalable Dense SLAM for Task-aware Optimization using CSFD, *ACM Trans. Graph.*, 2024, **43**(4), 1–15, DOI: 10.1145/3658233.
192 J. Lin, Z. Li, X. Tang, J. Liu, S. Liu, J. Liu, Y. Lu, X. Wu, S. Xu, Y. Yan and W. Yang, VastGaussian: Vast 3D Gaussians for Large Scene Reconstruction, *arXiv*, 2024, preprint, arXiv:2402.17427, DOI: 10.48550/arXiv.2402.17427.
193 E. Weber, A. Hołyński, V. Jampani, S. Saxena, N. Snavely, A. Kar and A. Kanazawa, NeRFiller: Completing Scenes via Generative 3D Inpainting, *arXiv*, 2023, preprint, arXiv:2312.04560, DOI: 10.48550/arXiv.2312.04560.
194 J. Seo, K. Fukuda, T. Shibuya, T. Narihira, N. Murata, S. Hu, C.-H. Lai, S. Kim and Y. Mitsufuji, GenWarp: Single Image to Novel Views with Semantic-Preserving Generative Warping, *arXiv*, 2024, preprint, arXiv:2405.17251, DOI: 10.48550/arXiv.2405.17251.
195 W. AI, *Introducing PRISM-1: Photorealistic reconstruction in static and dynamic scenes*, 2024, https://wayve.ai/thinking/prism-1/, accessed: 2024-06-17.
196 *3D-Aware Manipulation with Object-Centric Gaussian Splatting*, 2024, https://object-aware-gaussian.github.io/, accessed: 2024-06-17.
197 *Physically Embodied Gaussian Splatting: A Realtime Correctable World Model for Robotics*, 2024, https://embodied-gaussians.github.io/, accessed: 2024-06-17.
198 Y. Li and D. Pathak, *ICRA 2024 Workshop on 3D Visual Representations for Robot Manipulation*, 2024.
199 S. Xue, J. Dill, P. Mathur, F. Dellaert, P. Tsiotras and D. Xu, Neural Visibility Field for Uncertainty-Driven Active Mapping, *arXiv*, 2024, preprint, arXiv:2406.06948, DOI: 10.48550/arXiv.2406.06948.
200 N. R. Smalheiser, *J. Am. Soc. Inf. Sci. Technol.*, 2012, **63**, 218–224.
201 S. Henry and B. T. McInnes, *J. Biomed. Inf.*, 2017, **74**, 20–32.
202 M. Thilakaratne, K. Falkner and T. Atapattu, *ACM Comput. Surv.*, 2019, **52**, 1–34.
203 M. Krenn, R. Pollice, S. Y. Guo, M. Aldeghi, A. Cervera-Lierta, P. Friederich, G. dos Passos Gomes, F. Häse, A. Jinich, A. Nigam, Z. Yao and A. Aspuru-Guzik, *Nat. Rev. Phys.*, 2022, **4**, 761–769.
204 S. R. Young, A. Maksov, M. Ziatdinov, Y. Cao, M. Burch, J. Balachandran, L. Li, S. Somnath, R. M. Patton, S. V. Kalinin and R. K. Vasudevan, *J. Appl. Phys.*, 2018, **123**, 115303.
205 R. Kumar, A. Joshi, S. A. Khan and S. Misra, *Digital Discovery*, 2024, **3**, 944–953.
206 Y. Chiang, E. Hsieh, C.-H. Chou and J. Riebesell, LLaMP: Large Language Model Made Powerful for High-fidelity Materials Knowledge Retrieval and Distillation, *arXiv*, 2024, preprint, arXiv:2401.17244, DOI: 10.48550/arXiv.2401.17244.
207 M. C. Ramos, S. S. Michtavy, M. D. Porosoff and A. D. White, Bayesian Optimization of Catalysts With In-context Learning, *arXiv*, 2023, preprint, arXiv:2304.05341, DOI: 10.48550/arXiv.2304.05341.
208 B. M. Lake and M. Baroni, *Nature*, 2023, **623**, 115–121.
209 W. Liang, Y. Zhang, H. Cao, B. Wang, D. Ding, X. Yang, K. Vodrahalli, S. He, D. Smith, Y. Yin, D. McFarland and J. Zou, Can large language models provide useful feedback on research papers? A large-scale empirical analysis, *arXiv*, 2023, preprint, arXiv:2310.01783, DOI: 10.48550/arXiv.2310.01783.
210 Z. Qin, R. Jagerman, K. Hui, H. Zhuang, J. Wu, J. Shen, T. Liu, J. Liu, D. Metzler, X. Wang and M. Bendersky, Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting, *arXiv*, 2023, preprint, arXiv:2306.17563, DOI: 10.48550/arXiv.2306.17563.
211 J. Evans, J. D'Souza and S. Auer, Large Language Models as Evaluators for Scientific Synthesis, *arXiv*, 2024, preprint, arXiv:2407.02977, DOI: 10.48550/arXiv.2407.02977.
212 J. Fu, S.-K. Ng, Z. Jiang and P. Liu, GPTScore: Evaluate as You Desire, *arXiv*, 2023, arXiv:2302.04166, DOI: 10.48550/arXiv.2302.04166.
213 D. Paranyushkin, *The World Wide Web Conference*, New York, NY, USA, 2019, pp. 3584–3589.
214 M. Krenn, L. Buffoni, B. Coutinho, S. Eppel, J. G. Foster, A. Gritsevskiy, H. Lee, Y. Lu, J. P. Moutinho, N. Sanjabi, R. Sonthalia, N. M. Tran, F. Valente, Y. Xie, R. Yu and M. Kopp, *Nat. Mach. Intell.*, 2023, **5**, 1326–1335.
215 K. Yang, Y. Tian, N. Peng and D. Klein, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, 2022, pp. 4393–4479.
216 T. N. Fitria, *J. Engl. Lang. Teach.*, 2023, **12**(1), 44–58, DOI: 10.15294/elt.v12i1.64069.
217 J. V. Pavlik, *Journal. Mass Commun. Educat.*, 2023, **78**, 84–93.
218 S. Altmäe, A. Sola-Leyva and A. Salumets, *Reprod. Biomed. Online*, 2023, **47**, 3–9.
219 C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune and D. Ha, The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery, *arXiv*, 2024, preprint, arXiv:2408.06292, DOI: 10.48550/arXiv.2408.06292.
220 Y. Shao, Y. Jiang, T. A. Kanell, P. Xu, O. Khattab and M. S. Lam, Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models, *arXiv*, 2024, preprint, arXiv:2402.14207, DOI: 10.48550/arXiv.2402.14207.
221 S. Kambhampati, K. Valmeekam, L. Guan, M. Verma, K. Stechly, S. Bhambri, L. Saldyt and A. Murthy, LLMs Can't Plan, But Can Help Planning in LLM-Modulo

Frameworks, *arXiv*, 2024, preprint, arXiv:2402.01817, DOI: 10.48550/arXiv.2402.01817.

222 S. Farquhar, J. Kossen, L. Kuhn and Y. Gal, *Nature*, 2024, **630**, 625–630.

223 A. T. Kalai and S. S. Vempala, Calibrated Language Models Must Hallucinate, *arXiv*, 2024, preprint, arXiv:2311.14648, DOI: 10.48550/arXiv.2311.14648.

224 B. Mohammadi, Creativity Has Left the Chat: The Price of Debiasing Language Models, *arXiv*, 2024, arXiv:2406.05587, DOI: 10.48550/arXiv.2406.05587.

225 P. Sui, E. Duede, S. Wu and R. J. So, Confabulation: The Surprising Value of Large Language Model Hallucinations, *arXiv*, 2024, preprint, arXiv:2406.04175, DOI: 10.48550/arXiv.2406.04175.

226 M. Koivisto and S. Grassini, *Sci. Rep.*, 2023, **13**, 13601.

227 J. Haase and P. H. P. Hanel, Artificial muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity, *arXiv*, 2023, preprint, arXiv:2303.12003, DOI: 10.1016/j.yjoc.2023.100066.

228 K. Girotra, L. Meincke, C. Terwiesch and K. T. Ulrich, *SSRN Electron. J.*, 2023, **12**(1), 44–58.

229 L. Boussioux, J. N. Lane, M. Zhang, V. Jacimovic and K. R. Lakhani, *Harvard Business School Technology & Operations Mgt*. Unit Working Paper, 2023.

230 A. R. Doshi and O. Hauser, *SSRN*, 2023, DOI: 10.2139/ssrn.4535536.

231 B. S. Manning, K. Zhu and J. J. Horton, Automated Social Science: Language Models as Scientist and Subjects, *arXiv*, 2024, preprint, arXiv:2404.11794, DOI: 10.48550/arXiv.2404.11794.

232 Y. J. Ma, W. Liang, H.-J. Wang, S. Wang, Y. Zhu, L. Fan, O. Bastani and D. Jayaraman, DrEureka: Language Model Guided Sim-To-Real Transfer, *arXiv*, 2024, preprint, arXiv:2406.01967, DOI: 10.48550/arXiv.2406.01967.

233 Q. Wang, D. Downey, H. Ji and T. Hope, Learning to Generate Novel Scientific Directions with Contextualized Literature-based Discovery, *arXiv*, 2023, preprint, arXiv:2305.14259, DOI: 10.48550/arXiv.2305.14259.

234 Q. Wang, D. Downey, H. Ji and T. Hope, SciMON: Scientific Inspiration Machines Optimized for Novelty, *arXiv*, 2024, preprint, arXiv:2305.14259, DOI: 10.48550/arXiv.2305.14259.

235 C. Olah, A. Mordvintsev and L. Schubert, *Distill*, 2017, DOI: 10.23915/distill.00007, https://distill.pub/2017/feature-visualization.

236 R. Hendel, M. Geva and A. Globerson, In-Context Learning Creates Task Vectors, *arXiv*, 2023, preprint, arXiv:2310.15916, DOI: 10.48550/arXiv.2310.15916.

237 E. Todd, M. L. Li, A. S. Sharma, A. Mueller, B. C. Wallace and D. Bau, Function Vectors in Large Language Models, *arXiv*, 2024, preprint, arXiv:2310.15213, DOI: 10.48550/arXiv.2310.15213.

238 A. Arditi, O. Obeso, Aaquib111, wesg and N. Nanda, *Refusal in LLMs is mediated by a single direction, LessWrong*, 2024, https://www.lesswrong.com/posts/jGuXSZgv6qfdhMCuJ/refusal-in-llms-is-mediated-by-a-single-direction, accessed: 2024-06-11.

239 A. Zou, L. Phan, J. Wang, D. Duenas, M. Lin, M. Andriushchenko, R. Wang, Z. Kolter, M. Fredrikson and D. Hendrycks, Improving Alignment and Robustness with Circuit Breakers, *arXiv*, 2024, preprint, arXiv:2406.04313, DOI: 10.48550/arXiv.2406.04313.

240 K. Park, Y. J. Choe and V. Veitch, The Linear Representation Hypothesis and the Geometry of Large Language Models, *arXiv*, 2023, preprint, arXiv:2311.03658, DOI: 10.48550/arXiv.2311.03658.

241 K. Park, Y. J. Choe, Y. Jiang and V. Veitch, The Geometry of Categorical and Hierarchical Concepts in Large Language Models, *arXiv*, 2024, preprint, arXiv:2406.01506, DOI: 10.48550/arXiv.2406.01506.

242 T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, Z. Hatfield-Dodds, A. Tamkin, K. Nguyen, B. McLean, J. E. Burke, T. Hume, S. Carter, T. Henighan and C. Olah, *Transformer Circuits Thread*, 2023.

243 A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah and T. Henighan, *Transformer Circuits Thread*, 2024.

244 L. Gao, T. D. la Tour, H. Tillman, G. Goh, R. Troll, A. Radford, I. Sutskever, J. Leike and J. Wu, Scaling and evaluating sparse autoencoders, *arXiv*, 2024, preprint, arXiv:2406.04093, DOI: 10.48550/arXiv.2406.04093.

245 Y. Wang, W. Zhong, L. Li, F. Mi, X. Zeng, W. Huang, L. Shang, X. Jiang and Q. Liu, Aligning Large Language Models with Human: A Survey, *arXiv*, 2023, preprint, arXiv:2307.12966, DOI: 10.48550/arXiv.2307.12966.

246 E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang and W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, *arXiv*, 2021, preprint, arXiv:2106.09685, DOI: 10.48550/arXiv.2106.09685.

247 T. Dettmers, A. Pagnoni, A. Holtzman and L. Zettlemoyer, QLoRA: Efficient Finetuning of Quantized LLMs, *arXiv*, 2023, preprint, arXiv:2305.14314, DOI: 10.48550/arXiv.2305.14314.

248 S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng and M.-H. Chen, DoRA: Weight-Decomposed Low-Rank Adaptation, *arXiv*, 2024, preprint, arXiv:2402.09353, DOI: 10.48550/arXiv.2402.09353.

249 Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown and J. Kaplan, Constitutional A. I.: Harmlessness from AI Feedback,

arXiv, 2022, preprint, arXiv:2212.08073, DOI: **10.48550/arXiv.2212.08073**.

250 F. Song, B. Yu, M. Li, H. Yu, F. Huang, Y. Li and H. Wang, Preference Ranking Optimization for Human Alignment, *arXiv*, 2024, preprint, arXiv:2306.17492, DOI: **10.48550/arXiv.2306.17492**.

251 X. Li, P. Yu, C. Zhou, T. Schick, O. Levy, L. Zettlemoyer, J. Weston and M. Lewis, Self-Alignment with Instruction Backtranslation, *arXiv*, 2024, preprint, arXiv:2308.06259, DOI: **10.48550/arXiv.2308.06259**.

252 Z. Sun, Y. Shen, Q. Zhou, H. Zhang, Z. Chen, D. Cox, Y. Yang and C. Gan, *Principle-Driven Self-Alignment of Language Models from Scratch with Minimal Human Supervision*, 2023.

253 J. Pfau, A. Infanger, A. Sheshadri, A. Panda, J. Michael and C. Huebner, *Socially Responsible Language Modelling Research*, 2023.

254 N. Belrose, D. Schneider-Joseph, S. Ravfogel, R. Cotterell, E. Raff and S. Biderman, LEACE: Perfect linear concept erasure in closed form, *arXiv*, 2023, preprint, arXiv:2306.03819, DOI: **10.48550/arXiv.2306.03819**.

255 N. Belrose, Z. Furman, L. Smith, D. Halawi, I. Ostrovsky, L. McKinney, S. Biderman and J. Steinhardt, Eliciting Latent Predictions from Transformers with the Tuned Lens, *arXiv*, 2023, preprint, arXiv:2303.08112, DOI: **10.48550/arXiv.2303.08112**.

256 L. Aschenbrenner, *Situational Awareness: The Decade Ahead*, 2024, **https://situational-awareness.ai/wp-content/uploads/2024/06/situationalawareness.pdf**, accessed: 2024-06-07.

257 D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Nature*, 1986, **323**, 533–536.

258 M. Yuksekgonul, F. Bianchi, J. Boen, S. Liu, Z. Huang, C. Guestrin and J. Zou, TextGrad: Automatic "Differentiation" via Text, *arXiv*, 2024, preprint, arXiv:2406.07496, DOI: **10.48550/arXiv.2406.07496**.

259 W. Zhou, Y. Ou, S. Ding, L. Li, J. Wu, T. Wang, J. Chen, S. Wang, X. Xu, N. Zhang, H. Chen and Y. E. Jiang, Symbolic Learning Enables Self-Evolving Agents, *arXiv*, 2024, preprint, arXiv:2406.18532, DOI: **10.48550/arXiv.2406.18532**.

260 J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le and D. Zhou, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, *arXiv*, 2023, preprint, arXiv:2201.11903, DOI: **10.48550/arXiv.2201.11903**.

261 T. Kojima, S. S. Gu, M. Reid, Y. Matsuo and Y. Iwasawa, Large Language Models are Zero-Shot Reasoners, *arXiv*, 2023, preprint, arXiv:2205.11916, DOI: **10.48550/arXiv.2205.11916**.

262 J. Pfau, W. Merrill and S. R. Bowman, Let's Think Dot by Dot: Hidden Computation in Transformer Language Models, *arXiv*, 2024, preprint, arXiv:2404.15758, DOI: **10.48550/arXiv.2404.15758**.

263 L. Yang, Z. Yu, T. Zhang, S. Cao, M. Xu, W. Zhang, J. E. Gonzalez and B. Cui, Buffer of Thoughts: Thought-Augmented Reasoning with Large Language Models, *arXiv*, 2024, preprint, arXiv:2406.04271, DOI: **10.48550/arXiv.2406.04271**.

264 W. Chen, Y. Su, J. Zuo, C. Yang, C. Yuan, C.-M. Chan, H. Yu, Y. Lu, Y.-H. Hung, C. Qian, Y. Qin, X. Cong, R. Xie, Z. Liu, M. Sun and J. Zhou, AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors, *arXiv*, 2023, preprint, arXiv:2308.10848, DOI: **10.48550/arXiv.2308.10848**.

265 S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, C. Zhang, J. Wang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, C. Wu and J. Schmidhuber, MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework, *arXiv*, 2023, preprint, arXiv:2308.00352, DOI: **10.48550/arXiv.2308.00352**.

266 J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang and M. S. Bernstein, Generative Agents: Interactive Simulacra of Human Behavior, *arXiv*, 2023, preprint, arXiv:2304.03442, DOI: **10.48550/arXiv.2304.03442**.

267 M. Zhuge, H. Liu, F. Faccio, D. R. Ashley, R. Csordás, A. Gopalakrishnan, A. Hamdi, H. A. A. K. Hammoud, V. Herrmann, K. Irie, L. Kirsch, B. Li, G. Li, S. Liu, J. Mai, P. Piękos, A. Ramesh, I. Schlag, W. Shi, A. Stanić, W. Wang, Y. Wang, M. Xu, D.-P. Fan, B. Ghanem and J. Schmidhuber, Mindstorms in Natural Language-Based Societies of Mind, *arXiv*, 2023, preprint, arXiv:2305.17066, DOI: **10.48550/arXiv.2305.17066**.

268 I. Frisch and M. Giulianelli, LLM Agents in Interaction: Measuring Personality Consistency and Linguistic Alignment in Interacting Populations of Large Language Models, *arXiv*, 2024, preprint, arXiv:2402.02896, DOI: **10.48550/arXiv.2402.02896**.

269 T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest and X. Zhang, Large Language Model based Multi-Agents: A Survey of Progress and Challenges, *arXiv*, 2024, preprint, arXiv:2402.01680, DOI: **10.48550/arXiv.2402.01680**.

270 J. Wang, J. Wang, B. Athiwaratkun, C. Zhang and J. Zou, Mixture-of-Agents Enhances Large Language Model Capabilities, *arXiv*, 2024, preprint, arXiv:2406.04692, DOI: **10.48550/arXiv.2406.04692**.

271 Z. Wang, S. Cai, G. Chen, A. Liu, X. Ma and Y. Liang, Describe, Explain, Plan and Select: Interactive Planning with Large Language Models Enables Open-World Multi-Task Agents, *arXiv*, 2023, preprint, arXiv:2302.01560, DOI: **10.48550/arXiv.2302.01560**.

272 S. Abdelnabi, A. Gomaa, S. Sivaprasad, L. Schönherr and M. Fritz, LLM-Deliberation: Evaluating LLMs with Interactive Multi-Agent Negotiation Games, *arXiv*, 2023, preprint, arXiv:2309.17234, DOI: **10.48550/arXiv.2309.17234**.

273 Y. Dong, X. Jiang, Z. Jin and G. Li, Self-collaboration Code Generation via ChatGPT, *arXiv*, 2024, preprint, arXiv:2304.07590, DOI: **10.48550/arXiv.2304.07590**.

274 M. Wu, Y. Yuan, G. Haffari and L. Wang, (Perhaps) Beyond Human Translation: Harnessing Multi-Agent Collaboration for Translating Ultra-Long Literary Texts, *arXiv*, 2024, preprint, arXiv:2405.11804, DOI: **10.48550/arXiv.2405.11804**.

275 A. Bhoopchand, B. Brownfield, A. Collister, A. Dal Lago, A. Edwards, R. Everett, A. Frechette, Y. G. Oliveira, E. Hughes, K. W. Mathewson, P. Mendolicchio, J. Pawar, M. Pislar, A. Platonov, E. Senter, S. Singh, A. Zacherl and L. M. Zhang, *Nat. Commun.*, 2023, **14**, 7536.
276 J. Perez, C. Léger, M. Ovando-Tellez, C. Foulon, J. Dussauld, P.-Y. Oudeyer and C. Moulin-Frier, Cultural evolution in populations of Large Language Models, *arXiv*, 2024, preprint, arXiv:2403.08882, DOI: **10.48550/arXiv.2403.08882**.
277 A. L. Jones, Scaling Scaling Laws with Board Games, *arXiv*, 2021, preprint, arXiv:2104.03113, DOI: **10.48550/arXiv.2104.03113**.
278 R. Agarwal, A. Singh, L. M. Zhang, B. Bohnet, L. Rosias, S. Chan, B. Zhang, A. Anand, Z. Abbas, A. Nova, J. D. Co-Reyes, E. Chu, F. Behbahani, A. Faust and H. Larochelle, Many-Shot In-Context Learning, *arXiv*, 2024, preprint, arXiv:2404.11018, DOI: **10.48550/arXiv.2404.11018**.
279 R. Greenblatt, *Getting 50% (SoTA) on ARC-AGI with GPT-4o, Redwood Research blog on Substack*, 2024, **https://redwoodresearch.substack.com/p/getting-50-sota-on-arc-agi-with-gpt**, accessed: 2024-06-19.
280 M. Hassid, T. Remez, J. Gehring, R. Schwartz and Y. Adi, The Larger the Better? Improved LLM Code-Generation via Budget Reallocation, *arXiv*, 2024, preprint, arXiv:2404.00725, DOI: **10.48550/arXiv.2404.00725**.
281 B. Brown, J. Juravsky, R. Ehrlich, R. Clark, Q. V. Le, C. Ré and A. Mirhoseini, Large Language Monkeys: Scaling Inference Compute with Repeated Sampling, *arXiv*, 2024, preprint, arXiv:2407.21787, DOI: **10.48550/arXiv.2407.21787**.
282 C. Snell, J. Lee, K. Xu and A. Kumar, Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters, *arXiv*, 2024, preprint, arXiv:2408.03314, DOI: **10.48550/arXiv.2408.03314**.
283 Y. Wu, Z. Sun, S. Li, S. Welleck and Y. Yang, An Empirical Analysis of Compute-Optimal Inference for Problem-Solving with Language Models, *arXiv*, 2024, preprint, arXiv:2408.00724, DOI: **10.48550/arXiv.2408.00724**.
284 *Judgment Under Uncertainty: Heuristics and Biases*, ed. D. Kahneman, P. Slovic and A. Tversky, Cambridge University Press, Cambridge, 1982.
285 K. E. Stanovich and R. F. West, *Behav. Brain Sci.*, 2000, **23**, 645–665.
286 D. Kahneman, *Am. Psychol.*, 2003, **58**, 697–720.
287 J. S. Evans, *Trends Cognit. Sci.*, 2003, **7**, 454–459.
288 U. N. Sio and T. C. Ormerod, *Psychol. Bull.*, 2009, **135**, 94–120.
289 R. E. Beaty, M. Benedek, S. Barry Kaufman and P. J. Silvia, *Sci. Rep.*, 2015, **5**, 10964.
290 J. E. Driskell, R. P. Willis and C. Copper, *J. Appl. Psychol.*, 1992, **77**, 615–622.
291 A. Maravita and A. Iriki, *Trends Cognit. Sci.*, 2004, **8**, 79–86.
292 M. Csikszentmihalyi, *FLOW: The Psychology of Optimal Experience*, Harper and Row, 1990.
293 J. E. V. Gary, D. Ellis and C. Morris, *J. Leisure Res.*, 1994, **26**, 337–356.
294 J. Gold and J. Ciorciari, *Behav. Sci.*, 2020, **10**, 137.
295 J. Michael, S. Mahdi, D. Rein, J. Petty, J. Dirani, V. Padmakumar and S. R. Bowman, Debate Helps Supervise Unreliable Experts, *arXiv*, 2023, preprint, arXiv:2311.08702, DOI: **10.48550/arXiv.2311.08702**.
296 A. Khan, J. Hughes, D. Valentine, L. Ruis, K. Sachan, A. Radhakrishnan, E. Grefenstette, S. R. Bowman, T. Rocktäschel and E. Perez, Debating with More Persuasive LLMs Leads to More Truthful Answers, *arXiv*, 2024, preprint, arXiv:2402.06782, DOI: **10.48550/arXiv.2402.06782**.
297 J. W. A. Strachan, D. Albergo, G. Borghini, O. Pansardi, E. Scaliti, S. Gupta, K. Saxena, A. Rufo, S. Panzeri, G. Manzi, M. S. A. Graziano and C. Becchio, *Nat. Human Behav.*, 2024, **8**(7), 1285–1295, DOI: **10.1038/s41562-024-01882-z**.
298 W. Street, J. O. Siy, G. Keeling, A. Baranes, B. Barnett, M. McKibben, T. Kanyere, A. Lentz, B. A. y Arcas and R. I. M. Dunbar, LLMs achieve adult human performance on higher-order theory of mind tasks, *arXiv*, 2024, arXiv:2405.18870, DOI: **10.48550/arXiv.2405.18870**.
299 J. Connolly, F. Poli, P. Nugent, W. J. Shaw and K. G. Yager, *National Labs Should Be World-Leaders in Data Management*, Oppenheimer Science & Energy Leadership Program Think Pieces, 2021, **https://img1.wsimg.com/blobby/go/d0d92f6d-20cb-4140-aa26-dbe2979e28a1/downloads/OSELPCohort_4_Think-Piece_Summaries_2021.pdf**, accessed: 2024-06-12.
300 M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao and B. Mons, *Sci. Data*, 2016, **3**, 160018.
301 H. Lai, X. Liu, I. L. Iong, S. Yao, Y. Chen, P. Shen, H. Yu, H. Zhang, X. Zhang, Y. Dong and J. Tang, AutoWebGLM: Bootstrap And Reinforce A Large Language Model-based Web Navigating Agent, *arXiv*, 2024, preprint, arXiv:2404.03648, DOI: **10.48550/arXiv.2404.03648**.
302 J. Pan, Y. Zhang, N. Tomlin, Y. Zhou, S. Levine and A. Suhr, Autonomous Evaluation and Refinement of Digital Agents, *arXiv*, 2024, preprint, arXiv:2404.06474, DOI: **10.48550/arXiv.2404.06474**.
303 T. Xie, D. Zhang, J. Chen, X. Li, S. Zhao, R. Cao, T. J. Hua, Z. Cheng, D. Shin, F. Lei, Y. Liu, Y. Xu, S. Zhou, S. Savarese, C. Xiong, V. Zhong and T. Yu, OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments, *arXiv*, 2024, preprint, arXiv:2404.07972, DOI: **10.48550/arXiv.2404.07972**.

304 H. Bai, Y. Zhou, M. Cemri, J. Pan, A. Suhr, S. Levine and A. Kumar, DigiRL: Training In-The-Wild Device-Control Agents with Autonomous Reinforcement Learning, *arXiv*, 2024, preprint, arXiv:2406.11896, DOI: **10.48550/arXiv.2406.11896**.

305 Z. Zhao, T. Chavez, E. A. Holman, G. Hao, A. Green, H. Krishnan, D. McReynolds, R. J. Pandolfi, E. J. Roberts, P. H. Zwart, H. Yanxon, N. Schwarz, S. Sankaranarayanan, S. V. Kalinin, A. Mehta, S. I. Campbell and A. Hexemer, *4th Annual Workshop on Extreme-scale Experiment-in-the-Loop Computing*, XLOOP, 2022, pp. 10–15.

306 E. Zhang, V. Zhu, N. Saphra, A. Kleiman, B. L. Edelman, M. Tambe, S. M. Kakade and E. Malach, Transcendence: Generative Models Can Outperform The Experts That Train Them, *arXiv*, 2024, preprint, arXiv:2406.11741v1, DOI: **10.48550/arXiv.2406.11741**.