

Supplementary material for “AI4Research: a survey of artificial intelligence for scientific research”

1 Extended review of AI applications in scientific research

This appendix provides detailed surveys and technical descriptions corresponding to the high-level overview presented in Section 4 of the main text. It covers domain-specific methodologies, representative systems, and extended literature across natural sciences, applied sciences and engineering, and social sciences/humanities.

1.1 AI for natural science research

1.1.1 AI for physics research

In physics research, AI is now indispensable for developing new methodologies and driving discoveries [1, 2]. Its applications range from automated law discovery to physical world simulation and neural operator learning, all aimed at improving simulation accuracy, speeding up computation, and revealing hidden patterns from limited data [3, 4, 5].

Physical world simulation integrates physical priors with AI models to simulate complex systems while enforcing consistency with physical laws [6, 7, 8]. Earlier, Physics-Informed Neural Networks (PINNs) [9] embed PDE constraints in the loss function, allowing them to solve and infer nonlinear equations from sparse data. By exploiting the conserved-quantity structure of Hamiltonian mechanics, Hamiltonian Neural Networks [10] exploit conserved-quantity structures to enforce energy conservation, yielding faster convergence and drift-free, reversible simulations. Lagrangian Neural Networks [11] parameterize the system’s Lagrangian directly, avoiding coordinate choices, while still preserving exact energy conservation in examples like the double pendulum.

Automated law discovery leverages the reasoning power of LLMs, automated law discovery systems generate, test, and refine physical laws from noisy experimental data [12, 13]. For instance, AI-Newton [14] autonomously derives and validates physical laws, such as Newton’s laws and conservation principles, without requiring operator-provided equations. By integrating a solid knowledge base with a structured discovery workflow, AI-Newton generates interpretable models of physical phenomena. Shojaee *et al.* [12] propose a novel method utilizing LLMs’ scientific knowledge and code-generation capabilities to discover scientific equations directly from data. The DrSR framework [15] enhances law discovery by analyzing structural data relationships and implementing a feedback mechanism, improving performance across various domains. LLM-Feynman [16] combines automated feature engineering, LLM-based symbolic regression, and formula interpretation to extract interpretable expressions from both empirical data and domain knowledge. Recently, Li *et al.* [17] use enhanced visual prompting with domain expertise to uncover physical coordinates and governing equations from high-dimensional datasets more efficiently.

1.1.2 AI for biology & medical research

Artificial intelligence in the life sciences and medical research uses algorithms and computational models to analyze and predict across scales [18, 19, 20, 21, 22], from molecular structures to clinical diagnostics [23, 24, 25], to accelerate drug discovery [26, 27, 28], optimize experimental workflows [29, 30, 31], improve diagnostic accuracy, and advance precision medicine [32, 33, 34, 35].

Protein discovery. A notable example of computational innovation is Protein Discovery and protein structure prediction, which aims to predict the three-dimensional atomic structure of proteins. This is key for understanding biological functions and guiding drug design [36, 37, 38, 39]. For instance, Senior *et al.* [40] show that deep learning-based distance predictions significantly enhance de novo folding accuracy. The AlphaFold 2 system, developed by Jumper *et al.* [41], achieves atomic-level precision and has transformed structural biology since 2021. AlphaFold 3 [42] builds on this by introducing a diffusion-model

architecture that predicts monomeric structures and reconstructs protein-nucleic acid and protein-ligand complexes with near-experimental accuracy. Additionally, Lin *et al.* [43] present a dual-task LLaMA-based framework that integrates reaction and retrosynthesis into a unified recombination-fragmentation process, generating novel compounds with strong predicted protein-binding affinities through molecular docking feedback.

Cell & gene modeling. A crucial area of research is cell-level modeling and gene expression analysis, aiming to simulate cellular behavior and identify activity changes under various conditions [44, 45, 46, 47]. Several studies focus on pretraining models to improve cell or gene modeling [48, 49, 50]. However, due to the scarcity of high-quality gene and cellular data, recent work has explored data augmentation techniques to enhance AI training data [51, 52, 53]. Additionally, Roohani *et al.* [54] introduce an agent-based intelligent system that designs novel experiments, reasons about outcomes, and efficiently navigates hypothesis space by utilizing external tools to search the biomedical literature, analyze datasets, and engage secondary agents for evaluation, thus converging on optimal solutions. Furthermore, recent research has investigated AI-driven autonomous medical procedures, positioning AI as a collaborative tool for researchers [55, 56]. The micro-STAR system [57] integrates real-time OCT imaging with AI tissue classification to autonomously perform vascular suturing on ex vivo vessels, achieving leak-pressure performance comparable to expert surgeons, thus demonstrating the potential of AI and robotics in minimally invasive surgery.

Drug discovery In recent years, artificial intelligence (AI) has made significant advancements in the field of drug discovery, driving multi-faceted innovations in drug design and showcasing immense potential and prospects [26, 58, 59]. **(1) Structural Prediction and Molecular Design:** AI has made notable progress in structural prediction and molecular design. Early studies [60] use deep learning models to screen 23 potential antibiotic candidates from over 107 million molecules, successfully identifying a drug with antimicrobial activity. The LUMI-lab platform [61], which integrates molecular models with automated experimentation, discovering ionized lipids that excel in mRNA delivery. However, challenges related to data scarcity persist in AI-driven drug discovery. To mitigate this, strategies such as multi-target drug polypharmacology, decoding drug responses, and quantum computing have been proposed to enhance model performance [62]. **(2) Multi-Agent Collaborative Drug Identification:** Multi-agent systems have proven highly effective in drug discovery, facilitating the rapid identification of new therapeutic compounds [63]. For instance, the DrugAgent [64] and DrugPilot [65] automate machine learning programming through multi-agent collaboration, achieving full automation from data acquisition to model evaluation, thus improving efficiency. Solovev *et al.* [66] introduces multi-agent approach that combines LLMs with specialized generative models and validation tools to automate the end-to-end drug discovery process. Lee *et al.* [63] develop a multi-agent framework that retrieves and integrates information from biomedical knowledge bases to generate responses, avoiding the need for expensive domain-specific fine-tuning. Despite these advances, challenges remain in addressing data quality, model interpretability, and regulatory hurdles [27, 67]. **(3) Drug Repurposing:** Drug repurposing involves the use of approved drugs for new therapeutic indications [68, 69]. In liver fibrosis research, the AI-assisted Collaborative Scientist system has successfully recommended drugs with significant anti-fibrotic activity using human liver organoid platforms, showing promise for treating liver fibrosis [28]. By integrating knowledge graphs and diverse data sources, Liu *et al.* [64] and Gharizadeh *et al.* [70] identify potential drug repurposing candidates and provide interpretable predictions. Lee *et al.* [71] combine subgroup analysis and treatment effect estimation, simulating clinical trials to identify drug candidates and characterize patient subgroups based on treatment effects. This approach, tested on a real-world database of more than 8 million patients, simulate over 1,000 drug trials, identifying 14 drug candidates beneficial to specific subgroups.

Clinical diagnosis Clinical diagnosis advances through three converging breakthroughs. **(1) Clinical Brains:** First, LLMs serve as clinical “brains”, matching physician-level performance on medical licensing examinations [23, 72, 73, 74, 75]. Further, to enhance transparency and structure in decision-making, a human-AI note-taking framework [24] and Long-CoT reasoning techniques [76] has been proposed, leveraging case-based reasoning to guide clinical inquiry. **(2) Multi-Agent Hospital Simulation:** Multi-agent systems such as Agent Hospital replicate AI-AI [77, 78] and human-AI [79] collaborative

diagnostic and treatment workflows, effectively serving as an organizational “nervous system” for care coordination [80, 81]. **(3) Interactive Physical Actuation:** Robotic platforms guided by LLMs perform precise physical interventions. For example, an autonomous optical coherence tomography system delivers surgeon-level accuracy in delicate procedures such as vascular anastomosis [57]. Together, these breakthroughs demonstrate the feasibility of fully autonomous medical facilities in which artificial agents seamlessly integrate diagnostic reasoning, therapeutic planning, and procedural execution.

1.1.3 AI for chemistry & materials research

AI-driven automation in chemistry [82, 83, 84] and materials [85, 86, 87] integrates machine learning, robotics, and instrumentation into a closed-loop system for design, synthesis, and characterization, speeding decisions and experiments [88, 89, 90].

Automatic analysis seeks to identify optimal or novel material compositions in virtual or automated experimental setups while minimizing the number of required experiments [91, 92, 93]. Specifically, Chen *et al.* [94] introduce MEGNet, demonstrating that graph neural networks can achieve density functional theory-level accuracy for both molecular and crystalline properties. Li *et al.* [95] employ two-stage Bayesian optimization to screen 560 organic photocatalysts using only 2.4% of the experiment conditions, thereby obtaining significantly improved performance. Ekosso *et al.* [96] combine low-cost robotic platforms with high-throughput microscopy and Gaussian process models to map vesicle formation processes. More recently, Szymanski *et al.* [97] implement a robot-machine-learning platform that accelerate compound discovery and identify 41 novel inorganic materials within 17 days.

Automatic discovery is an automated experimental platform that combines robotic operations, online characterization, and real-time decision-making algorithms to autonomously execute the full experimental process, from reagent dispensing to result analysis [98, 99, 100]. Early research laid both theoretical and practical foundations: Butler *et al.* [98] review machine learning methods that accelerate materials design and discovery [101, 102, 103]. Furthermore, Dai *et al.* [104] employ a mobile robot with UPLC-MS and NMR to plan and interpret syntheses in a manner akin to human chemists. Jayarathna *et al.* [105] leverage literature data to reduce the number of experiments in an active-learning loop, discovering new Ru-based catalysts. Dai *et al.* [106] introduce an AI advisor for ion-electron polymers, improving performance by 150% over spin-coating methods. More recently, several studies have incorporated LLMs to enhance innovation and knowledge in chemistry and materials discovery [107, 108, 109].

Full human-AI collaboration process management leverages LLMs or natural-language understanding and generation to support hypothesis formulation, experimental design, and iterative optimization, aiming to facilitate more intuitive and efficient research interactions [110, 111, 112]. The AILA framework [113] embeds LLMs within a fully automated atomic-force microscopy workflow, illustrating both the potential and current constraints of language models in guiding real-time microscopic experiments. Sprueill *et al.* [114] and Feng *et al.* [115] integrate LLM-driven linguistic reasoning with chemical feedback in a heuristic search loop to propose novel catalysts and reaction pathways within uncertain chemical spaces. Recognizing that collective intelligence often outperforms individual reasoning, several studies [116, 117, 118] employ multi-agent architectures in which LLMs generate hypotheses, design experiments and direct iterative optimization, thereby achieving seamless human-AI collaboration. Ma *et al.* [119] introduce the first fully automated retrosynthetic planning agent tailored for LLM-driven macromolecular design, enabling comprehensive enumeration of viable multi-branch reaction routes. Meanwhile, Zhu *et al.* [120] demonstrate a robotic AI chemist that performs ore pretreatment and catalyst optimization on Martian meteorite samples.

1.2 AI for applied science and engineering research

1.2.1 AI for robotics and control research

AI for Robotics and Control Research applies AI methods: deep learning, reinforcement learning, and large language models, to the perception, decision-making, and control of robots, aiming to boost adaptability, robustness, and autonomy in novel environments [121, 122].

Autonomous design & optimization systems integrate robotics, machine learning, and domain expertise to automate experiment planning, execution, and optimization [99]. Uddin *et al.* [123] introduce OptoMate, a system using a fine-tuned language model for optical setup design and a robotic arm to assemble spectroscopy components with submillimeter precision, enabling cloud-based optical labs. Meiszczyński *et al.* [124] employ computer vision and feedforward neural networks in a feedback optimizer to adjust 3D printing parameters in real-time, reducing data collection from days to hours and ensuring consistent part quality. Angello *et al.* [125] apply physically informed feature selection and supervised learning in a closed-loop system to enhance photostability and uncover solvent-mediated triplet-state mechanisms. Bu *et al.* [126] combine a text-conditioned video diffusion model with a feedback-driven controller to generate visual plans and iteratively refine actions, significantly improving performance.

End-to-end vision-based control End-to-end vision-based control feeds raw images or video frames directly into a neural network to generate control signals, removing separate perception, planning, and control modules. Early work combine convolutional neural networks (CNNs) with reinforcement learning or guided policy search to map camera inputs to motion commands. Levine *et al.* [127] first apply Guided Policy Search to train perception and control jointly, mapping raw images to motor torques and demonstrating reliable real-world grasping. Levine *et al.* [128] train a CNN on massive real grasp attempts to predict grasp success in real time, closing the loop on novel objects. Kalashnikov *et al.* [129] introduce a self-supervised, closed-loop Q-learning framework trained on 580,000 grasps, enabling dynamic strategy adjustment, retries, and disturbance resilience. Tobin *et al.* [130] propose domain randomization, varying simulator rendering parameters so models trained on synthetic data transfer directly to real-world detection and grasping.

Sim-to-real robustness & safety ensures reliable transfer of simulation-trained policies to real-world tasks while adhering to safety constraints. Bochem *et al.* [131] integrate sharpness-aware optimization into gradient-based RL to identify flat minima, enhancing transfer robustness in contact-rich tasks without compromising sample efficiency. Ayabe *et al.* [132] assess offline RL methods on a legged robot subjected to random and adversarial torque disturbances, revealing vulnerability to sudden perturbations and emphasizing the need for real-time adaptation and safety measures. Radosavovic *et al.* [133] train a Transformer-based controller using deep RL and deploy it outdoors for one week without safety scaffolding, showcasing adaptive performance amidst disturbances, rugged terrain, and varying payloads. Yang *et al.* [134] apply domain randomization for vision-based servoing of soft robots, eliminating the need for on-robot fine-tuning and enabling direct transfer of simulation-trained models to continuum manipulators. Guerrier *et al.* [135] combine control barrier functions with RL to enforce safety constraints during learning, preventing hazardous states in complex environments.

Multi-task & multi-agent control frameworks facilitate concurrent task execution or enable collaboration among agents in complex workflows, thereby enhancing parallelism and automation. Tahmid and Notomista [136] introduce a reinforcement learning-based framework designed to dynamically learn and compose task policies in robotic systems with redundant architectures, incorporating time-varying priority stacks to adjust task priorities. Team *et al.* [137] propose a unified multi-agent system capable of automatically generating hypotheses, designing and conducting experiments, and refining methods through iterative feedback, establishing a closed-loop process that accelerates interdisciplinary research.

1.2.2 AI for software engineering

AI for Software Engineering Research focuses on applying AI techniques to automate software development tasks, enhance code quality, and improve developer productivity. This includes code generation, bug detection, code review, and software testing.

Code generation refers to the use of AI models to automatically generate code snippets or entire programs from natural language descriptions or existing code patterns [138, 139, 140, 141]. This can accelerate the development process and reduce manual coding [142, 143]. For instance, Chen *et al.* [144] develop Codex, a GPT model fine-tuned on GitHub’s publicly available code, which supports GitHub

Copilot. To democratize program synthesis, Nijkamp *et al.* [145] train and release CodeGen, an LLM based on both natural and programming language data, alongside the open-source training library JAXFORMER. Additionally, several studies have explored advanced code capabilities and support for multiple programming languages, like Python, Java, and R [138, 139, 140].

End-to-end software development covers the entire software development lifecycle, with AI automating various stages [146, 147, 148, 149]. For example, Phan *et al.* [150] develop HyperAgent, a generalist multi-agent system designed to handle various SE tasks across different programming languages, mimicking human developers' workflows. Qian *et al.* [151] introduce Experiential Co-Learning, which enables software development agents to leverage historical experiences to improve task performance. Meanwhile, Qian *et al.* [152] introduce ChatDev, a chat-based framework for software development, and Kang *et al.* [153] present an explainable automated debugging framework powered by LLM-driven scientific debugging.

1.3 AI for humanities and social science research

AI has been widely utilized to automate the design, execution, and analysis of humanities and social science experiments, encompassing tasks from hypothesis generation to data collection, with minimal human intervention. In this context, we will focus on two key domains:

1.3.1 AI for sociology research

AI in sociology research refers to the use of machine learning, natural language processing, and multi-agent systems to simulate, analyze, and explore social phenomena [154, 155, 156]. Through AI, researchers can reconstruct macro-level patterns of collective behavior and gain deeper insights into micro-level cultural contexts and individual interactions, thereby revitalizing traditional sociological methods [157].

AI-assisted experimental and interview studies. Controlled experiments and simulated interviews are increasingly employed by scholars to test social science hypotheses and evaluate the effects of various social mechanisms and policy interventions. Manning *et al.* [158] propose a methodology that combines structural causal models with large language models to automatically generate and empirically validate social science hypotheses in contexts such as negotiations, bail hearings, job interviews, and auctions. This approach effectively bridges the gap between theory and practice by utilizing the model both as a scientific tool for hypothesis generation and as an experimental subject for validation. Liu and Yu [159] develop MimiTalk, an automated interview system, and conduct a comparative study of AI-led and human-led interviews with 20 participants on the Prolific platform. This study demonstrates the feasibility of AI-mediated interviews and highlights their potential in experimental settings.

Large-scale simulation of social phenomena. This approach leverages algorithmic tools to automate the collection and analysis of extensive textual, visual, and interaction data to simulate and examine the macro-level dynamics of community practices and value evolution [160, 161, 162]. Perez *et al.* [163] automate the extraction and analysis of large-scale text and image datasets to map cultural practices, value systems, and trends in contemporary online communities [164, 165]. Zamudio *et al.* [166] propose a simulation framework for cultural evolution using multi-agent LLMs, enabling the manipulation of network structures, individual traits, and biases in information transmission to investigate factors driving cultural diffusion and change. Chen *et al.* [167] develop a GPT-based three-module framework, including information extraction, variant generation, and outcome prediction, that achieved high consistency in predicting outcomes across 319 economic field experiments, while also reflecting the impact of gender, race, and social norms on performance. Additionally, Bao *et al.* [168] reveal the underlying, often unspoken codes within societies.

Potential risks discussion. While LLMs demonstrate strong predictive capabilities in the natural sciences, their performance in the social sciences remains limited. Manning *et al.* [158] find that, although LLMs can predict the signs of estimated effects well when given a proposed structural causal model, they struggle to predict the magnitudes reliably. Additionally, Luke *et al.* [169] highlight that LLMs face challenges in handling treatment effect heterogeneity and exhibit systematic biases when predicting

social science outcomes. As such, LLMs' predictive capabilities are still underdeveloped, particularly in forecasting novel empirical patterns that could inform future experimentation [170].

1.3.2 AI for psychology research

Research methodology focuses on the design, implementation, and validation of psychological experiments to ensure validity and reproducibility [171, 172, 173].

Experiment workflow automation and simulation. Recent research has explored integrating AI into the management and data simulation of psychology experiments [174, 175, 176]. Zamudio *et al.* [166] introduce the RAISE pipeline, automating the generation and validation of visual stimuli. In five experiments, AI-generated images match researcher-designed stimuli in both validity and recognizability. Cingillioglu *et al.* [177] conduct a fully automated online RCT with 1,193 participants, where AI managed recruitment, random assignment, intervention delivery, and data collection, successfully replicating eight classical hypotheses with gold-standard rigor. Cui *et al.* [178], Strachan *et al.* [179], Suri *et al.* [180] use GPT-4 to simulate responses for 154 classical experiments, reproducing 76% of primary effects but yielding 71.6% unexpected significant outcomes, illustrating the promise of AI-assisted replication while emphasizing the need for cautious interpretation [181, 182, 183].

Human-AI trust and safety design. Research on human-AI trust explores the development of trust during human-AI interactions and derives strategies for ensuring safety [184]. Li *et al.* [185] introduce a three-dimensional framework encompassing the trustor, trustee, and context, identifying key factors influencing trust and offering design recommendations for improving safety. Building on this, Chandra *et al.* [186], through interviews with 283 individuals who have mental health experiences, develop a taxonomy comprising 19 risky AI behaviors and 21 negative psychological impacts. From this, they propose a multi-path case-method framework and a set of safety guidelines aimed at mitigating these risks.

Psychological interventions. Psychological interventions increasingly employ AI-driven chatbots to provide scalable and cost-effective psychological support [187]. Earlier, Hagendorff *et al.* [188], Dillion *et al.* [182], and Binz and Schulz [176] discuss whether and when LLMs can replace human participants in psychological research, reviewing early evidence and proposing a theoretical framework while highlighting methodological caveats. In a randomized controlled trial, Heinz *et al.* [189] find that the Therabot chatbot resulted in significant reductions in clinical-level symptoms compared to the control group. Similarly, Spytska *et al.* [190] explore the use of the Friend chatbot for crisis support, demonstrating that its efficacy is comparable to traditional face-to-face therapy. These findings highlight the potential of generative AI to enhance accessibility and effectiveness in mental health services, particularly in settings with limited resources [191, 192]. Recent work further demonstrates that LLMs can match or exceed human performance in generating emotionally resonant narratives [193, 194] and even pass standard Turing tests [195], underscoring their broader psychological and communicative capabilities.

1.3.3 AI for archaeology

Research methodology in archaeology is currently undergoing a fundamental transformation, shifting from a discipline reliant on manual, labor-intensive data collection and qualitative interpretation to a computational model capable of processing vast, heterogeneous datasets [196]. This paradigm shift, often described as the move toward "post-digital" archaeology, focuses on the design, implementation, and validation of machine learning (ML) workflows to ensure validity, reproducibility, and scalability in the analysis of material culture [197].

Automated landscape archaeology. AI has become increasingly central to automated landscape archaeology, enabling the detection of anthropogenic features across satellite, aerial, and LiDAR data [198, 199]. CNN-based methods, often supported by transfer learning and data augmentation, have proven effective for large-scale feature detection [200], while LiDAR visualization techniques combined with deep networks enhance the recognition of terrain-based structures under dense vegetation [201]. More advanced approaches such as Mask R-CNN [202] allow pixel-level mapping of architectural remains, providing finer

structural detail than traditional bounding-box detection. Despite these advances, models trained on idealized examples still struggle to generalize to real-world variability, underscoring the continued necessity of human-in-the-loop verification [203].

Artifact analysis. With the help of artificial intelligence, artifact analysis has been increasingly transformed by automating the classification of ceramics, lithics, and osteological materials, and by unlocking information embedded in legacy archaeological documents [204, 205]. CNN-based systems such as those developed in the ArchAIDE project apply image recognition and typological learning, often enhanced with standardized archaeological drawings, to classify fragmented or eroded objects more reliably [206, 207]. Beyond classification, AI pipelines like AutArch use neural object detection to extract artifact images and associated metadata (e.g., scale bars, north arrows) directly from heterogeneous printed or PDF sources, enabling the rapid construction of unified datasets that previously required extensive manual labor [208]. Together, these advances demonstrate how AI can both democratize specialist typology work and reactivate centuries of archival excavation data for large-scale archaeological analysis.

Artifacts restoration. Recent advances in cultural heritage restoration increasingly rely on computational methods that combine computer vision, language modeling, and robotic automation to recover and reconstruct damaged or incomplete artifacts. In textual restoration, volumetric ink detection, exemplified by the Vesuvius Challenge, uses machine learning to analyze CT scans of carbonized papyri, enabling virtual unwrapping and decipherment of previously unreadable texts [209], while Transformer-based models such as DeepMind’s Ithaca restore missing characters and contextual information in ancient inscriptions [210]. In parallel, physical artifact reconstruction has been enhanced by systems like the RePAIR workflow, which integrates computer vision with robotic arms, applying genetic algorithms and geometric matching to align fragments based on geometry and texture, and autonomously manipulating them into place [211]. Together, these approaches illustrate how AI-driven methods, often in human–AI collaborative workflows, can significantly improve both textual and material reconstruction in archaeology, extending the reach of expert interpretation and enabling tasks that were previously infeasible.

1.3.4 AI for humanities and arts research

Research in Digital Humanities and Arts leverages AI to transition from qualitative hermeneutics to quantitative modeling and generative simulation, enabling scalable analysis of cultural artifacts and experimental validation of aesthetic theories. AI is increasingly employed in cultural and creative studies, combining computational modeling, large language models, and generative systems to analyze, quantify, and augment human cultural and artistic activity. In **quantitative cultural analysis**, Bayesian hierarchical models and LLM-based simulations are used to disentangle factors shaping literary style and emotional expression and to model the transmission and evolution of narratives across generations, revealing patterns of cultural drift and cumulative cultural evolution under different social network structures [163, 212]. In **creative generation** domains, generative AI influences both human cognition and aesthetic perception [213, 214]: collaborative studies show that AI enhances novelty for novices while improving quality and refinement for experts, acting as a “skill amplifier,” [215] and psychological experiments indicate that AI-generated artworks, optimized for symmetry and color harmony, are often indistinguishable from human-made art, demonstrating that generative models capture high-level artistic regularities [216]. Together, these applications illustrate how AI can rigorously quantify abstract cultural phenomena, augment human creative performance, and shape aesthetic judgments.

References

- 1 Dorigo T, Giammanco A, Vischia P, et al. Toward the end-to-end optimization of particle physics instruments with differentiable programming. *Reviews in Physics*, 2023, 10: 100085
- 2 Boehnlein A, Diefenthaler M, Sato N, et al. Colloquium: Machine learning in nuclear physics. *Reviews of modern physics*, 2022, 94: 031003

- 3 Jiao L, Song X, You C, et al. Ai meets physics: A comprehensive survey. *Artificial Intelligence Review*, 2024, 57: 256
- 4 Wang Y, Bai J, Lin Z, et al. Artificial intelligence for partial differential equations in computational mechanics: A review. *arXiv preprint arXiv:2410.19843*, 2024
- 5 Meng C, Griesemer S, Cao D, et al. When physics meets machine learning: A survey of physics-informed machine learning. *Machine Learning for Computational Science and Engineering*, 2025, 1: 1–23
- 6 Battaglia P, Pascanu R, Lai M, et al. Interaction networks for learning about objects, relations and physics. *Advances in Neural Information Processing Systems*, 2016, 29
- 7 de Avila Belbute-Peres F, Smith K, Allen K, et al. End-to-end differentiable physics for learning and control. *Advances in Neural Information Processing Systems*, 2018, 31
- 8 Raissi M, Perdikaris P, Ahmadi N, et al. Physics-informed neural networks and extensions. *arXiv preprint arXiv:2408.16806*, 2024
- 9 Raissi M, Perdikaris P, Karniadakis G E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 2019, 378: 686–707
- 10 Greydanus S, Dzamba M, Yosinski J. Hamiltonian neural networks. *Advances in Neural Information Processing Systems*, 2019, 32
- 11 Cranmer M, Greydanus S, Hoyer S, et al. Lagrangian neural networks. *arXiv preprint arXiv:2003.04630*, 2020
- 12 Shojaee P, Meidani K, Gupta S, et al. LLM-SR: Scientific equation discovery via programming with large language models. In: *The Thirteenth International Conference on Learning Representations*, 2025
- 13 Shojaee P, Nguyen N H, Meidani K, et al. LLM-SRBench: A new benchmark for scientific equation discovery with large language models. In: *Forty-second International Conference on Machine Learning*, 2025
- 14 Fang Y L, Jian D S, Li X, et al. Ai-newton: A concept-driven physical law discovery system without prior physical knowledge. *arXiv preprint arXiv:2504.01538*, 2025
- 15 Wang R, Wang B, Li K, et al. Drsr: Llm based scientific equation discovery with dual reasoning from data and experience. *arXiv preprint arXiv:2506.04282*, 2025
- 16 Song Z, Ju M, Ren C, et al. Llm-feynman: Leveraging large language models for universal scientific formula and theory discovery. *arXiv preprint arXiv:2503.06512*, 2025
- 17 Li R, Lu Y, Tang S, et al. Mllm-based discovery of intrinsic coordinates and governing equations from high-dimensional data. *arXiv preprint arXiv:2505.11940*, 2025
- 18 Lam H Y I, Ong X E, Mutwil M. Large language models in plant biology. *Trends in Plant Science*, 2024
- 19 Li W, Mao Z, Xiao Z, et al. Large language model for knowledge synthesis and ai-enhanced biomanufacturing. *Trends in Biotechnology*, 2025
- 20 Yang L, Xu S, Sellergren A, et al. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*, 2024
- 21 Wang H, He Y, Coelho P P, et al. Spatialagent: An autonomous ai agent for spatial biology. *bioRxiv*, 2025: 2025–04

- 22 Buess L, Keicher M, Navab N, et al. From large language models to multimodal ai: A scoping review on the potential of generative ai in medicine. *arXiv preprint arXiv:2502.09242*, 2025
- 23 Wu C, Lei J, Zheng Q, et al. Can gpt-4v (ision) serve medical applications? Case studies on gpt-4v for multimodal medical diagnosis. *arXiv preprint arXiv:2310.09909*, 2023
- 24 Craig D B. A human-LLM note-taking system with case-based reasoning as framework for scientific discovery. In: *Proceedings of the 1st Workshop on AI and Scientific Discovery: Directions and Opportunities*, 2025. 22–30
- 25 Wang W, Ma Z, Wang Z, et al. A survey of llm-based agents in medicine: How far are we from baymax? *arXiv preprint arXiv:2502.11211*, 2025
- 26 Tang X, Dai H, Knight E, et al. A survey of generative ai for de novo drug design: New frontiers in molecule and protein generation. *Briefings in Bioinformatics*, 2024, 25: bbae338
- 27 Othman Z K, Ahmed M M, Okesanya O J, et al. Advancing drug discovery and development through gpt models: A review on challenges, innovations and future prospects. *Intelligence-Based Medicine*, 2025: 100233
- 28 Guan Y, Inchai J, Fang Z, et al. Ai-assisted drug re-purposing for human liver fibrosis. *bioRxiv*, 2025: 2025–04
- 29 Swanson K, Wu W, Bulaong N L, et al. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, 2025: 1–3
- 30 Ifargan T, Hafner L, Kern M, et al. Autonomous llm-driven research—from data to human-verifiable research papers. *NEJM AI*, 2025, 2: AIoa2400555
- 31 Newsham I, Kovačević L, Moulange R, et al. Large language models for zero-shot inference of causal structures in biology. *arXiv preprint arXiv:2503.04347*, 2025
- 32 Jiang X, Ou L, Chen Y, et al. A fuzzy logic-based approach to predict human interaction by functional near-infrared spectroscopy. *IEEE Transactions on Fuzzy Systems*, 2025
- 33 Kapitonova M, Ball T. Human-ai teaming using large language models: Boosting brain-computer interfacing (bci) and brain research. *arXiv preprint arXiv:2501.01451*, 2024
- 34 Yang J J, Hwang S H. Transforming hematological research documentation with large language models: An approach to scientific writing and data analysis. *Blood Research*, 2025, 60: 1–11
- 35 Huang K, Zhang S, Wang H, et al. Biomni: A general-purpose biomedical ai agent. *bioRxiv*, 2025: 2025–05
- 36 Ding N, Qu S, Xie L, et al. Automating exploratory proteomics research via language models. *arXiv preprint arXiv:2411.03743*, 2024
- 37 Ghafarollahi A, Buehler M J. Protagents: Protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery*, 2024, 3: 1389–1409
- 38 Ghafarollahi A, Buehler M J. Sparks: Multi-agent artificial intelligence model discovers protein design principles. *arXiv preprint arXiv:2504.19017*, 2025
- 39 Pei Q, Wu L, Gao K, et al. Leveraging biomolecule and natural language through multi-modal learning: A survey. *arXiv preprint arXiv:2403.01528*, 2024
- 40 Senior A W, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 2020, 577: 706–710
- 41 Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 2021, 596: 583–589

- 42 Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 2024, 630: 493–500
- 43 Lin X, Liu Q, Xiang H, et al. Enhancing chemical reaction and retrosynthesis prediction with large language model and dual-task learning. *arXiv preprint arXiv:2505.02639*, 2025
- 44 Bunne C, Roohani Y, Rosen Y, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 2024, 187: 7045–7063
- 45 Garijo D, Yang Q, Vargas H, et al. Neurodisk: An ai approach to automate continuous inquiry-driven discoveries in neuroimaging genetics. *bioRxiv*, 2025
- 46 Rood J E, Hupalowska A, Regev A. Toward a foundation model of causal cell and tissue biology with a perturbation cell and tissue atlas. *Cell*, 2024, 187: 4520–4545
- 47 Luo X, Rechart A, Sun G, et al. Large language models surpass human experts in predicting neuroscience results. *Nature Human Behaviour*, 2025, 9: 305–315
- 48 Chen Y, Zou J. Genept: A simple but effective foundation model for genes and cells built from chatgpt. *bioRxiv*, 2024: 2023–10
- 49 Bian H, Chen Y, Luo E, et al. General-purpose pre-trained large cellular models for single-cell transcriptomics. *National Science Review*, 2024, 11: nwae340
- 50 Consens M E, Dufault C, Wainberg M, et al. Transformers and genome language models. *Nature Machine Intelligence*, 2025: 1–17
- 51 Afonja T, Sheth I, Binkytė R, et al. Llm4grn: Discovering causal gene regulatory networks with llms—evaluation through synthetic data generation. *arXiv preprint arXiv:2410.15828*, 2024
- 52 Maleki S, Huetter J C, Chuang K V, et al. Efficient fine-tuning of single-cell foundation models enables zero-shot molecular perturbation prediction. *arXiv preprint arXiv:2412.13478*, 2024
- 53 Agraz M, Goksuluk D, Zhang P, et al. Ml-gap: Machine learning-enhanced genomic analysis pipeline using autoencoders and data augmentation. *Frontiers in Genetics*, 2024, 15: 1442759
- 54 Roohani Y, Lee A, Huang Q, et al. Biodiscoveryagent: An ai agent for designing genetic perturbation experiments. *arXiv preprint arXiv:2405.17631*, 2024
- 55 Zou J, Topol E J. The rise of agentic ai teammates in medicine. *The Lancet*, 2025, 405: 457
- 56 Xiao Y, Liu J, Zheng Y, et al. Cellagent: An llm-driven multi-agent framework for automated single-cell data analysis. *arXiv preprint arXiv:2407.09811*, 2024
- 57 Haworth J, Biswas R, Opfermann J, et al. Autonomous robotic system with optical coherence tomography guidance for vascular anastomosis. *arXiv preprint arXiv:2410.07493*, 2024
- 58 Gupta R, Srivastava D, Sahu M, et al. Artificial intelligence to deep learning: Machine intelligence approach for drug discovery. *Molecular diversity*, 2021, 25: 1315–1360
- 59 Edfeldt K, Edwards A M, Engkvist O, et al. A data science roadmap for open science organizations engaged in early-stage drug discovery. *Nature Communications*, 2024, 15: 5640
- 60 Stokes J M, Yang K, Swanson K, et al. A deep learning approach to antibiotic discovery. *Cell*, 2020, 180: 688–702
- 61 Cui H, Xu Y, Pang K, et al. Lumi-lab: A foundation model-driven autonomous platform enabling discovery of new ionizable lipid designs for mrna delivery. *BioRxiv*, 2025: 2025–02
- 62 Gangwal A, Ansari A, Ahmad I, et al. Current strategies to address data scarcity in artificial intelligence-based drug discovery: A comprehensive review. *Computers in Biology and Medicine*, 2024, 179: 108734

- 63 Lee N, De Brouwer E, Hajiramezanali E, et al. Rag-enhanced collaborative llm agents for drug discovery. arXiv preprint arXiv:2502.17506, 2025
- 64 Liu S, Lu Y, Chen S, et al. Drugagent: Automating ai-aided drug discovery programming through llm multi-agent collaboration. arXiv preprint arXiv:2411.15692, 2024
- 65 Li K, Wu Z, Wang S, et al. Drugpilot: Llm-based parameterized reasoning agent for drug discovery. arXiv preprint arXiv:2505.13940, 2025
- 66 Solovev G V, Zhidkovskaya A B, Orlova A, et al. Towards llm-driven multi-agent pipeline for drug discovery: Neurodegenerative diseases case study. In: 2nd AI4Research Workshop: Towards a Knowledge-grounded Scientific Research Lifecycle, 2024
- 67 Yuan S, Färber M. Hallucinations can improve large language models in drug discovery. arXiv preprint arXiv:2501.13824, 2025
- 68 Lu Y, Hu Y, Li C. Drugclip: Contrastive drug-disease interaction for drug repurposing. arXiv preprint arXiv:2407.02265, 2024
- 69 Huang K, Chandak P, Wang Q, et al. A foundation model for clinician-centered drug repurposing. *Nature Medicine*, 2024, 30: 3601–3613
- 70 Gharizadeh A, Abbasi K, Ghareyazi A, et al. Hgtdr: Advancing drug repurposing with heterogeneous graph transformers. *Bioinformatics*, 2024, 40: btae349
- 71 Lee S, Liu R, Cheng F, et al. A deep subgrouping framework for precision drug repurposing via emulating clinical trials on real-world patient data. arXiv preprint arXiv:2412.20373, 2024
- 72 Brodsky V, Ullah E, Bychkov A, et al. Generative artificial intelligence in anatomic pathology. *Archives of Pathology & Laboratory Medicine*, 2025
- 73 Gottweis J, Weng W H, Daryin A, et al. Towards an ai co-scientist. arXiv preprint arXiv:2502.18864, 2025
- 74 Khalifa M, Albadawy M, Iqbal U. Advancing clinical decision support: The role of artificial intelligence across six domains. *Computer Methods and Programs in Biomedicine Update*, 2024, 5: 100142
- 75 Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*, 2023, 620: 172–180. URL <https://doi.org/10.1038/s41586-023-06291-2>
- 76 Lan W, Wang W, Ji C, et al. Clinicalgpt-r1: Pushing reasoning capability of generalist disease diagnosis with large language model. arXiv preprint arXiv:2504.09421, 2025
- 77 Fan Z, Tang J, Chen W, et al. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. arXiv preprint arXiv:2402.09742, 2024
- 78 Kyung D, Chung H, Bae S, et al. Patientsim: A persona-driven simulator for realistic doctor-patient interactions. arXiv preprint arXiv:2505.17818, 2025
- 79 Sayin B, Schlicht I B, Hong N V, et al. Medsyn: Enhancing diagnostics with human-ai collaboration. arXiv preprint arXiv:2506.14774, 2025
- 80 Li J, Lai Y, Li W, et al. Agent hospital: A simulacrum of hospital with evolvable medical agents. arXiv preprint arXiv:2405.02957, 2024
- 81 Bao Z, Liu Q, Guo Y, et al. Piors: Personalized intelligent outpatient reception based on large language model with multi-agents medical scenario simulation. arXiv preprint arXiv:2411.13902, 2024

- 82 Hayes T, Rao R, Akin H, et al. Simulating 500 million years of evolution with a language model. *Science*, 2025: eads0018
- 83 M Bran A, Cox S, Schilter O, et al. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 2024, 6: 525–535
- 84 Mroz A M, Basford A R, Hastedt F, et al. Cross-disciplinary perspectives on the potential for artificial intelligence across chemistry. *Chemical Society Reviews*, 2025
- 85 Pyzer-Knapp E O, Pitera J W, Staar P W, et al. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Computational Materials*, 2022, 8: 84
- 86 Jiang X, Xue D, Wang W Y, et al. Ai4materials: Transforming the landscape of materials science and engineering. *Review of Materials Research*, 2025: 100010
- 87 Choi J, Lee B. Accelerating materials language processing with large language models. *Communications Materials*, 2024, 5: 13
- 88 Chen X, Buriak J M, Salanne M, et al. Nano & ai: A nobel partnership, 2024
- 89 Yuan W, Chen G, Wang Z, et al. Empowering generalist material intelligence with large language models. *Advanced Materials*, 2025: 2502771
- 90 Wu J, He J, Liu H, et al. From literature to lab: Hardware-independent autonomous chemical synthesis with reinforcement learning. In: *Companion Proceedings of the ACM on Web Conference 2025*, 2025. 2923–2926
- 91 Batista B C, Amrutha S, Yan J, et al. High-throughput robotic collection, imaging, and machine learning analysis of salt patterns: Composition and concentration from dried droplet photos. *Digital Discovery*, 2025, 4: 1030–1041
- 92 Rajabi-Kochi M, Mahboubi N, Gill A P S, et al. Adaptive representation of molecules and materials in bayesian optimization. *Chemical Science*, 2025, 16: 5464–5474
- 93 Seo J, Kim D, Jeong J, et al. FlavorDiffusion: Modeling food-chemical interactions with diffusion. In: *Proceedings of the 1st Workshop on AI and Scientific Discovery: Directions and Opportunities*, 2025. 70–77
- 94 Chen C, Ye W, Zuo Y, et al. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 2019, 31: 3564–3572
- 95 Li X, Che Y, Chen L, et al. Sequential closed-loop bayesian optimization as a guide for organic molecular metallophotocatalyst formulation discovery. *Nature Chemistry*, 2024, 16: 1286–1294
- 96 Ekosso C, Liu H, Glagovich A, et al. Accelerating the discovery of abiotic vesicles with ai-guided automated experimentation. *Langmuir*, 2024, 41: 858–867
- 97 Szymanski N J, Rendy B, Fei Y, et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 2023, 624: 86–91
- 98 Butler K T, Davies D W, Cartwright H, et al. Machine learning for molecular and materials science. *Nature*, 2018, 559: 547–555
- 99 Liang H, Wang C, Yu H, et al. Real-time experiment-theory closed-loop interaction for autonomous materials science. *arXiv preprint arXiv:2410.17430*, 2024
- 100 Darvish K, Skreta M, Zhao Y, et al. Organa: A robotic assistant for automated chemistry experimentation and characterization. *Matter*, 2025, 8
- 101 Merchant A, Batzner S, Schoenholz S S, et al. Scaling deep learning for materials discovery. *Nature*, 2023, 624: 80–85

- 102 Dangayach R, Jeong N, Demirel E, et al. Machine learning-aided inverse design and discovery of novel polymeric materials for membrane separation. *Environmental Science & Technology*, 2024, 59: 993–1012
- 103 Huang X, Zhao C, Wang H, et al. Ai-assisted inverse design of sequence-ordered high intrinsic thermal conductivity polymers. *Materials Today Physics*, 2024, 44: 101438
- 104 Dai T, Vijayakrishnan S, Szczypiński F T, et al. Autonomous mobile robots for exploratory synthetic chemistry. *Nature*, 2024: 1–8
- 105 Jayarathna R, Onsree T, Drummond S, et al. Experimental discovery of novel ammonia synthesis catalysts via active learning. *Journal of Materials Chemistry A*, 2024, 12: 3046–3060
- 106 Dai Y, Chan H, Vriza A, et al. Adaptive ai decision interface for autonomous electronic material discovery. *arXiv preprint arXiv:2504.13344*, 2025
- 107 Kristiadi A, Strieth-Kalthoff F, Skreta M, et al. A sober look at llms for material discovery: Are they actually good for bayesian optimization over molecules? *arXiv preprint arXiv:2402.05015*, 2024
- 108 Zhang X, Ni Z, Fanady B, et al. Chatgpt-assisted rational design for iterative performance optimization of perovskite solar cells. Available at SSRN 5127472
- 109 Yang Y, Shi R, Li Z, et al. Batgpt-chem: A foundation large model for chemical engineering. *Research*, 2025, 8: 0827
- 110 Wang H, Skreta M, Ser C T, et al. Efficient evolutionary search over chemical space with large language models. *arXiv preprint arXiv:2406.16976*, 2024
- 111 Luo M, Xie Z, Li H, et al. Physics-informed, dual-objective optimization of high-entropy-alloy nanozymes by a robotic ai chemist. *Matter*, 2025, 8
- 112 Leng Y, Zhong Y, Gu Z, et al. Intelligent, personalized scientific assistant via large language models for solid-state battery research. *ACS Materials Letters*, 2025, 7: 1807–1816
- 113 Mandal I, Soni J, Zaki M, et al. Autonomous microscopy experiments through large language model agents. *arXiv preprint arXiv:2501.10385*, 2024
- 114 Sprueill H W, Edwards C, Agarwal K, et al. Chemreasoner: Heuristic search over a large language model’s knowledge space using quantum-chemical feedback. *arXiv preprint arXiv:2402.10980*, 2024
- 115 Feng R, Liang Y, Yin T, et al. Agentic assistant for materials scientists. *The Electrochemical Society Interface*, 2025, 34: 45–48
- 116 Ni Z, Li Y, Hu K, et al. Matpilot: An llm-enabled ai materials scientist under the framework of human-machine collaboration. *arXiv preprint arXiv:2411.08063*, 2024
- 117 Song T, Luo M, Zhang X, et al. A multiagent-driven robotic ai chemist enabling autonomous chemical research on demand. *Journal of the American Chemical Society*, 2025, 147: 12534–12545
- 118 Lai Z, Pu Y. Prim: Principle-inspired material discovery through multi-agent collaboration. *arXiv preprint arXiv:2504.08810*, 2025
- 119 Ma Q, Zhou Y, Li J. Automated retrosynthesis planning of macromolecules using large language models and knowledge graphs. *Macromolecular Rapid Communications*, 2025: 2500065
- 120 Zhu Q, Huang Y, Zhou D, et al. Automated synthesis of oxygen-producing catalysts from martian meteorites by a robotic ai chemist. *Nature Synthesis*, 2024, 3: 319–328
- 121 Lee S K, Ko H. Generative machine learning in adaptive control of dynamic manufacturing processes: A review. *arXiv preprint arXiv:2505.00210*, 2025

- 122 Lange R T, Prasad A, Sun Q, et al. The ai cuda engineer: Agentic cuda kernel discovery, optimization and composition. Technical Report, Technical report, Sakana AI, 02 2025, 2025
- 123 Uddin S Z, Vaidya S, Choudhary S, et al. Ai-driven robotics for free-space optics. arXiv preprint arXiv:2505.17985, 2025
- 124 Mieszczanek P, Corke P, Mehanian C, et al. Towards industry-ready additive manufacturing: Ai-enabled closed-loop control for 3d melt electrowriting. *Communications Engineering*, 2024, 3: 158
- 125 Angello N H, Friday D M, Hwang C, et al. Closed-loop transfer enables artificial intelligence to yield chemical knowledge. *Nature*, 2024, 633: 351–358
- 126 Bu Q, Zeng J, Chen L, et al. Closed-loop visuomotor control with generative expectation for robotic manipulation. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024
- 127 Levine S, Finn C, Darrell T, et al. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 2016, 17: 1–40
- 128 Levine S, Pastor P, Krizhevsky A, et al. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International journal of robotics research*, 2018, 37: 421–436
- 129 Kalashnikov D, Irpan A, Pastor P, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In: *Conference on robot learning*, 2018. 651–673
- 130 Tobin J, Fong R, Ray A, et al. Domain randomization for transferring deep neural networks from simulation to the real world. In: *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, 2017. 23–30
- 131 Bochem S, Gonzalez-Sanchez E, Bicker Y, et al. Improving generalization of robot locomotion policies via sharpness-aware reinforcement learning. arXiv preprint arXiv:2411.19732, 2024
- 132 Ayabe S, Otomo T, Kera H, et al. Robustness evaluation of offline reinforcement learning for robot control against action perturbations. arXiv preprint arXiv:2412.18781, 2024
- 133 Radosavovic I, Xiao T, Zhang B, et al. Real-world humanoid locomotion with reinforcement learning. *Science Robotics*, 2024, 9: eadi9579
- 134 Yang H J, Khosravi M, Walt B, et al. Zero-shot sim-to-real transfer for reinforcement learning-based visual servoing of soft continuum arms. arXiv preprint arXiv:2504.16916, 2025
- 135 Guerrier M, Soma K, Fouad H, et al. Guided by guardrails: Control barrier functions as safety instructors for robotic learning. arXiv preprint arXiv:2505.18858, 2025
- 136 Tahmid S A, Notomista G. Value iteration for learning concurrently executable robotic control tasks. arXiv preprint arXiv:2504.01174, 2025
- 137 Team N, Zhang B, Feng S, et al. Novelseek: When agent becomes the scientist—building closed-loop system from hypothesis to verification. arXiv preprint arXiv:2505.16938, 2025
- 138 Roziere B, Gehring J, Gloeckle F, et al. Code llama: Open foundation models for code. arXiv preprint arXiv:2308.12950, 2023
- 139 Guo D, Zhu Q, Yang D, et al. Deepseek-coder: When the large language model meets programming—the rise of code intelligence. arXiv preprint arXiv:2401.14196, 2024
- 140 Li R, Allal L B, Zi Y, et al. Starcoder: May the source be with you! arXiv preprint arXiv:2305.06161, 2023

- 141 Huang J, Feng X, Chen Q, et al. Mldebugging: Towards benchmarking code debugging across multi-library scenarios. arXiv preprint arXiv:2506.13824, 2025
- 142 Lozhkov A, Li R, Allal L B, et al. Starcoder 2 and the stack v2: The next generation. arXiv preprint arXiv:2402.19173, 2024
- 143 Zhang Y, Su J, Sun Y, et al. Seed-coder: Let the code model curate data for itself. arXiv preprint arXiv:2506.03524, 2025
- 144 Chen M, Tworek J, Jun H, et al. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374, 2021
- 145 Nijkamp E, Pang B, Hayashi H, et al. Codegen: An open large language model for code with multi-turn program synthesis. arXiv preprint arXiv:2203.13474, 2022
- 146 Ozkaya I. Application of large language models to software engineering tasks: Opportunities, risks, and implications. *IEEE Software*, 2023, 40: 4–8
- 147 Fan A, Gokkaya B, Harman M, et al. Large language models for software engineering: Survey and open problems. In: 2023 IEEE/ACM International Conference on Software Engineering: Future of Software Engineering (ICSE-FoSE), 2023. 31–53
- 148 Jimenez C E, Yang J, Wettig A, et al. SWE-bench: Can language models resolve real-world github issues? In: The Twelfth International Conference on Learning Representations, 2024
- 149 Le Hai N, Nguyen D M, Bui N D. Repoexec: Evaluate code generation with a repository-level executable benchmark. arXiv preprint arXiv:2406.11927, 2024
- 150 Phan H N, Nguyen T N, Nguyen P X, et al. Hyperagent: Generalist software engineering agents to solve coding tasks at scale. arXiv preprint arXiv:2409.16299, 2024
- 151 Qian C, Dang Y, Li J, et al. Experiential co-learning of software-developing agents. arXiv preprint arXiv:2312.17025, 2023
- 152 Qian C, Liu W, Liu H, et al. Chatdev: Communicative agents for software development. arXiv preprint arXiv:2307.07924, 2023
- 153 Kang S, Chen B, Yoo S, et al. Explainable automated debugging via large language model-driven scientific debugging. *Empirical Software Engineering*, 2025, 30: 1–28
- 154 Loffredo J R, Yun S. Agent-enhanced large language models for researching political institutions. arXiv preprint arXiv:2503.13524, 2025
- 155 Xia Y, Qu A, Zheng Y, et al. Reimagining urban science: Scaling causal inference with large language models. arXiv preprint arXiv:2504.12345, 2025
- 156 Karjus A. Machine-assisted quantizing designs: Augmenting humanities and social sciences with artificial intelligence. *Humanities and Social Sciences Communications*, 2025, 12: 1–18
- 157 Li Z, Abramson C M. Ethnography and machine learning: Synergies and new directions. arXiv preprint arXiv:2412.06087, 2024
- 158 Manning B S, Zhu K, Horton J J. Automated social science: Language models as scientist and subjects. Technical Report, National Bureau of Economic Research, 2024
- 159 Liu F, Yu S. Step further towards automated social science: An ai-powered interview platform. Available at SSRN, 2024
- 160 Hoffmann M P, Fillies J, Paschke A. Malinowski in the age of ai: Can large language models create a text game based on an anthropological classic? arXiv preprint arXiv:2410.20536, 2024

- 161 Sreedhar K, Cai A, Ma J, et al. Simulating cooperative prosocial behavior with multi-agent llms: Evidence and mechanisms for ai agents to inform policy decisions. In: Proceedings of the 30th International Conference on Intelligent User Interfaces, 2025. 1272–1286
- 162 Huang Y, Wang X, Liu H, et al. Adasociety: An adaptive environment with social structures for multi-agent decision-making. arXiv preprint arXiv:2411.03865, 2024
- 163 Perez J, Léger C, Ovando-Tellez M, et al. Cultural evolution in populations of large language models. arXiv preprint arXiv:2403.08882, 2024
- 164 Yu H, Hong Z, Cheng Z, et al. Researchtown: Simulator of human research community. arXiv preprint arXiv:2412.17767, 2024
- 165 Sheldon Z, Kumar P. Economic anthropology in the era of generative artificial intelligence. arXiv preprint arXiv:2410.15238, 2024
- 166 Zamudio C, Grigsby J L, Michelsen M. Raise: A new method to develop experimental stimuli for advertising research with image generative artificial intelligence. *Journal of Advertising*, 2024: 1–16
- 167 Chen Y, Hu Y, Lu Y. Predicting field experiments with large language models. arXiv preprint arXiv:2504.01167, 2025
- 168 Bao H, Wu S, Choi J, et al. Language models surface the unwritten code of science and society. arXiv preprint arXiv:2505.18942, 2025
- 169 Luke H, Ashwini A, Isaias G, et al. Predicting results of social science experiments using large language models, 2024. URL <https://samim.io/dl/Predicting%20results%20of%20social%20science%20experiments%20using%20large%20language%20models.pdf>, predicting Results of Social Science Experiments Using Large Language Models
- 170 Lehr S A, Caliskan A, Liyanage S, et al. Chatgpt as research scientist: Probing gpt’s capabilities as a research librarian, research ethicist, data generator, and data predictor. *Proceedings of the National Academy of Sciences*, 2024, 121: e2404328121
- 171 Tong S, Mao K, Huang Z, et al. Automating psychological hypothesis generation with ai: When large language models meet causal graph. *Humanities and Social Sciences Communications*, 2024, 11: 1–14
- 172 Li B, Guan J, Dou L, et al. Can large language models understand you better? An mbti personality detection dataset aligned with population traits. arXiv preprint arXiv:2412.12510, 2024
- 173 Zhang G, Yu Z, Li N, et al. Human behavior simulation: Objectives, methodologies, and open problems. arXiv preprint arXiv:2412.07788, 2024
- 174 Qiu H, Lan Z. Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions. arXiv preprint arXiv:2408.15787, 2024
- 175 Wang J, Xiao Y, Li Y, et al. Towards a client-centered assessment of llm therapists by client simulation. arXiv preprint arXiv:2406.12266, 2024
- 176 Binz M, Schulz E. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 2023, 120. URL <http://dx.doi.org/10.1073/pnas.2218523120>
- 177 Cingillioglu I, Gal U, Prokhorov A. Ai-experiments in education: An ai-driven randomized controlled trial for higher education research. *Education and Information Technologies*, 2024, 29: 19649–19677
- 178 Cui Z, Li N, Zhou H. Can ai replace human subjects? A large-scale replication of psychological experiments with llms. arXiv preprint arXiv:2409.00128, 2024

- 179 Strachan J W A, Albergo D, Borghini G, et al. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 2024, 8: 1285–1295. URL <https://www.nature.com/articles/s41562-024-01895-1>
- 180 Suri G, Slater L R, Ziaee A, et al. Do large language models show decision heuristics similar to humans? A case study using gpt-3.5. *Journal of Experimental Psychology: General*, 2024, 153: 1066
- 181 Chen Y, Liu T X, Shan Y, et al. The emergence of economic rationality of gpt. *Proceedings of the National Academy of Sciences*, 2023, 120: e2316205120
- 182 Dillion D, Tandon N, Gu Y, et al. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 2023, 27: 597–600. URL <https://www.sciencedirect.com/science/article/pii/S1364661323000980>
- 183 Goli A, Singh A. Frontiers: Can large language models capture human preferences? *Marketing Science*, 2024, 43: 709–722. URL <https://doi.org/10.1287/mksc.2023.0306>
- 184 Qin L, Huang S, Chen Q, et al. Mmsd2. 0: Towards a reliable multi-modal sarcasm detection system. *arXiv preprint arXiv:2307.07135*, 2023
- 185 Li Y, Wu B, Huang Y, et al. Developing trustworthy artificial intelligence: Insights from research on interpersonal, human-automation, and human-ai trust. *Frontiers in Psychology*, 2024, 15: 1382693
- 186 Chandra M, Naik S, Ford D, et al. From lived experience to insight: Unpacking the psychological risks of using ai conversational agents. *arXiv preprint arXiv:2412.07951*, 2024
- 187 Praveen S, Gajjar P, Ray R K, et al. Crafting clarity: Leveraging large language models to decode consumer reviews. *Journal of Retailing and Consumer Services*, 2024, 81: 103975. URL <https://www.sciencedirect.com/science/article/pii/S0969698924002716>
- 188 Hagendorff T, Fabi S, Kosinski M. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 2023, 3: 833–838. URL <http://dx.doi.org/10.1038/s43588-023-00527-x>
- 189 Heinz M V, Mackin D M, Trudeau B M, et al. Randomized trial of a generative ai chatbot for mental health treatment. *Nejm Ai*, 2025, 2: AIoa2400802
- 190 Spytska L. The use of artificial intelligence in psychotherapy: Development of intelligent therapeutic systems. *BMC psychology*, 2025, 13: 175
- 191 de Kok T. Chatgpt for textual analysis? How to use generative llms in accounting research. *Management Science*, 2025. URL <https://doi.org/10.1287/mnsc.2023.03253>, published online in *Articles in Advance*, 13 Jan 2025
- 192 Deiner M S, Honcharov V, Li J, et al. Large language models can enable inductive thematic analysis of a social media corpus in a single prompt: Human validation study. *JMIR Infodemiology*, 2024, 4: e59641. URL <https://doi.org/10.2196/59641>
- 193 Wen Y, Laporte S. Experiential narratives in marketing: A comparison of generative ai and human content. *Journal of Public Policy & Marketing*, 2025, 44: 392–410. URL <https://doi.org/10.1177/07439156241297973>
- 194 Rousmaniere T, Li X, Zhang Y, et al. Large language models as mental health resources: Patterns of use in the united states, 2025
- 195 Jones C R, Bergen B K. Large language models pass the turing test. *arXiv preprint arXiv:2503.23674*, 2025
- 196 Gattiglia G. Managing artificial intelligence in archeology. an overview. *Journal of Cultural Heritage*, 2025, 71: 225–233

- 197 Dey A. The application of artificial intelligence in detecting archaeological elements: A comprehensive review. *The Academic*, 2025, 3
- 198 Kadhim I, Abed F M. A critical review of remote sensing approaches and deep learning techniques in archaeology. *Sensors*, 2023, 23: 2918
- 199 Zeynali R, Mandanici E, Bitelli G. A technical note on ai-driven archaeological object detection in airborne lidar derivative data, with cnn as the leading technique. *Remote Sensing*, 2025, 17: 2733
- 200 Zimmer-Dauphinee J, VanValkenburgh P, Wernke S A. Eyes of the machine: Ai-assisted satellite archaeological survey in the andes. *Antiquity*, 2024, 98: 245–259
- 201 Somrak M, Džeroski S, Kokalj Ž. Learning to classify structures in als-derived visualizations of ancient maya settlements with cnn. *Remote Sensing*, 2020, 12: 2215
- 202 Fylaktos P, Petropoulos G P, Lemesios I. A deep learning method for the automated mapping of archaeological structures from geospatial data: A case study of delos island. *ISPRS International Journal of Geo-Information*, 2025, 14: 220
- 203 Sobotkova A, Kristensen-McLachlan R D, Mallon O, et al. Validating predictions of burial mounds with field data: The promise and reality of machine learning. *Journal of Documentation*, 2024, 80: 1167–1189
- 204 Ling Z, Delnevo G, Salomoni P, et al. Findings on machine learning for identification of archaeological ceramics-a systematic literature review. *IEEE Access*, 2024
- 205 Mnasri Z, D’Andrea A. Automatic inventory of archaeological artifacts based on object detection and classification using deep and transfer learning. *Digital Applications in Archaeology and Cultural Heritage*, 2025: e00458
- 206 Anichini F, Dershowitz N, Dubbini N, et al. The automatic recognition of ceramics from only one photo: The archaide app. *Journal of Archaeological Science: Reports*, 2021, 36: 102788
- 207 Santos J, Nunes D A, Padnevyh R, et al. Automatic ceramic identification using machine learning. lusitanian amphorae and faience. two portuguese case studies. *STAR: Science & Technology of Archaeological Research*, 2024, 10: e2343214
- 208 Klein K, Muller A, Wohde A, et al. Autarch: An ai-assisted workflow for object detection and automated recording in archaeological catalogues. *arXiv preprint arXiv:2311.17978*, 2023
- 209 Parsons S, Parker C S, Chapman C, et al. Educelab-scrolls: Verifiable recovery of text from herculaneum papyri using x-ray ct. *arXiv preprint arXiv:2304.02084*, 2023
- 210 Assael Y, Sommerschield T, Shillingford B, et al. Restoring and attributing ancient texts using deep neural networks. *Nature*, 2022, 603: 280–283
- 211 Tsesmelis T, Palmieri L, Khoroshiltseva M, et al. Re-assembling the past: The repair dataset and benchmark for real world 2d and 3d puzzle solving. *Advances in Neural Information Processing Systems*, 2024, 37: 30076–30105
- 212 Konle L, Kröncke M, Winko S, et al. Connecting the dots. variables of literary history and emotions in german-language poetry. *Journal of Computational Literary Studies*, 2024, 2
- 213 Croitoru F A, Hondru V, Ionescu R T, et al. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2023, 45: 10850–10869
- 214 Zhang Y, Diddee H, Holm S, et al. Noveltybench: Evaluating language models for humanlike diversity. *arXiv preprint arXiv:2504.05228*, 2025
- 215 Wang N, Kim H, Peng J, et al. Exploring creativity in human–ai co-creation: A comparative study across design experience. *Frontiers in Computer Science*, 2025, 7: 1672735
- 216 Van Hees J, Grootswagers T, Quek G L, et al. Human perception of art in the age of artificial intelligence. *Frontiers in Psychology*, 2025, 15: 1497469