NEJM AI

DOI: 10.1056/AIoa2400555

ORIGINAL ARTICLE

Autonomous LLM-Driven Research — from Data to Human-Verifiable Research Papers

Tal Ifargan, BSc.,[1] Lukas Hafner, Ph.D.,[2] Maor Kern,[3] Ori Alcalay, BSc.,[3] and Roy Kishony, Ph.D.[2,4,5]

Received: June 4, 2024; Revised: October 10, 2024; Accepted: October 17, 2024; Published: December 3, 2024

BACKGROUND Artificial intelligence (AI) promises to accelerate scientific discovery, but it remains unclear whether AI systems can perform fully autonomous research, and whether they can do so while adhering to key scientific values, such as transparency, traceability, and verifiability. The aim of this study was to develop and evaluate an AI-automation platform that performs transparent, traceable, and human-verifiable scientific research.

METHODS To mimic human scientific practices, we built "data-to-paper," an automation platform that guides interacting large language model (LLM) agents through a complete stepwise research process that starts with annotated data and results in comprehensive research papers, while programmatically backtracing information flow and allowing human oversight and interactions. The platform can run fully autonomously (in autopilot mode) or with human intervention (in copilot mode).

RESULTS In autopilot mode, provided only with annotated data, data-to-paper raised hypotheses; designed research plans; wrote and debugged analysis codes; generated and interpreted results; and created complete, information-traceable research papers. Even though the research novelty of manuscripts created by data-to-paper was relatively limited, the process demonstrated the autonomous generation of de novo quantitative insights from data, such as unraveling associations between health indicators and clinical outcomes. For simple research goals and datasets, a fully autonomous cycle can create manuscripts that independently recapitulate the findings of peer-reviewed biomedical publications without major errors in about 80 to 90% of cases. Yet, as goal or data complexity increases, human copiloting becomes critical for ensuring accuracy and overall quality. By tracking information flow through the steps, the platform creates "data-chained" manuscripts, in which downstream results are programmatically linked to upstream code and data, thus setting a new standard for the verifiability of scientific outputs.

CONCLUSIONS Our work demonstrates the potential for AI-driven acceleration of scientific discovery in data-driven biomedical research and beyond, while enhancing, rather than jeopardizing, traceability, transparency, and verifiability.

Introduction

Advances in natural language processing have resulted in large language models (LLMs) such as ChatGPT (Chat Generative Pretrained Transformer) that are capable of writing text, answering questions, and generating code at a human level.[1-5]

Mr. Ifargan and Dr. Hafner contributed equally to this article.

The author affiliations are listed at the end of the article.

Dr. Kishony can be contacted at rkishony@technion.ac.il.

Augmenting LLMs with external tools, as well as with automated iterative algorithmic prompting and multiagent interactions, has enabled them to tackle more complex, multistep tasks, such as solving mathematical problems,[6-8] coding and debugging large code projects,[9,10] and creating book-long texts and scripts.[11] Most recently, LLMs have even demonstrated a capacity to design and run experiments as well as perform clinical diagnostics and evaluate scientific research.[12-15] Yet, despite these advances, scientific research, in particular the de novo creation of insights from data, remains a stronghold of human intelligence and ingenuity.[16-21] Limitations of LLMs in this regard impede a potential increase in scientific discovery, especially in data-rich fields such as biomedicine and epidemiology. The recent advancement of artificial intelligence (AI) has led to spirited discussions about the potential and risks of the technology's application in science[22] and to emerging guidelines emphasizing the importance of key values, including accountability, oversight, transparency, and verifiability, which are notoriously challenging in AI.[23]

Conducting research and compiling results and conclusions into a transparent and methodologically traceable and verifiable scientific paper is a highly challenging task, involving multiple interconnected steps and requiring planning, inference, and deduction, as well as meticulous tracing of information. Although scientists may, in principle, follow a myriad of creative paths toward discovery, certain conventional research paths have been established.[24] These conventional paths typically follow an almost canonical sequence of steps: formulating and reshaping a research question in light of the literature, designing and executing a research plan, interpreting the results in the context of prior knowledge, and writing a research paper. Beyond this well-established multistep structure, the human-driven scientific process has three additional key characteristics. First, the process is not linear; it often requires iteratively reverting to earlier steps. Second, it is built on rigorous tracing and control of both textual and quantitative information between steps. Third, at each of the steps, human scientists rely on feedback from peers, mentors, and external reviewers, enabling a collective expertise that extends beyond individual capabilities. Taken together, these key features make science a unique process of human creativity.

Inspired by the research practices of human scientists, we built "data-to-paper," an automation platform that systematically guides multiple LLM and rule-based algorithmic agents through the conventional steps of data-driven scientific research, with automated feedback, iterative cycles of review and revision, and structured control and tracing of information flow between these research steps. We specifically focused on a relatively simple and well-defined process of hypothesis-testing research on preexisting annotated data. Starting with such a dataset, the process was designed to raise hypotheses; write, debug, and execute code to analyze the data and perform statistical tests; interpret the results; and write well-structured scientific papers that not only describe results and conclusions, but transparently delineate the research methodologies, allowing human scientists to understand, repeat, and verify the analysis.

Discussions of emerging guidelines for AI-driven science[23] have served as a design framework for data-to-paper. The process of designing the system in light of these discussions has yielded a fully transparent, traceable, and verifiable workflow, and the algorithmic chaining of data, methodology, and results that allows for the tracing of downstream results all the way back to the specific parts of code that generated them. The system can run with or without a predefined research goal (in fixed-goal or open-goal modalities) and with or without human interactions and feedback (in copilot or autopilot modes).

We performed three open-goal and two fixed-goal case studies on different publicly available datasets[25-29] and evaluated the AI-driven research process as well as the novelty and accuracy of the scientific papers created by data-to-paper. We show that data-to-paper, running fully autonomously (in autopilot mode), can perform complete and correct run cycles for simple datasets and research goals. However, for complex research, human copiloting becomes critical.

Methods

To autonomously analyze a user-provided dataset and create a research paper, data-to-paper guides multiple LLM and rule-based agents through a series of predefined research steps, each of which is designed to create well-defined quantitative or textual research products (Fig. 1). The process includes the following steps: data exploration; literature search and iterative formulation of a research goal and hypothesis; creation of a hypothesis-testing plan; writing of data analysis code; creation of scientific tables; searching for related literature; and writing of the paper section by section (Figs. 1A and 1B, top; 17 steps in total). The research goal can also be provided as human input, in which case the goal-determining steps are skipped (fixed-goal modality, dashed bypass arrow, Fig. 1A; Supplementary Methods in Supplementary Appendix 1).

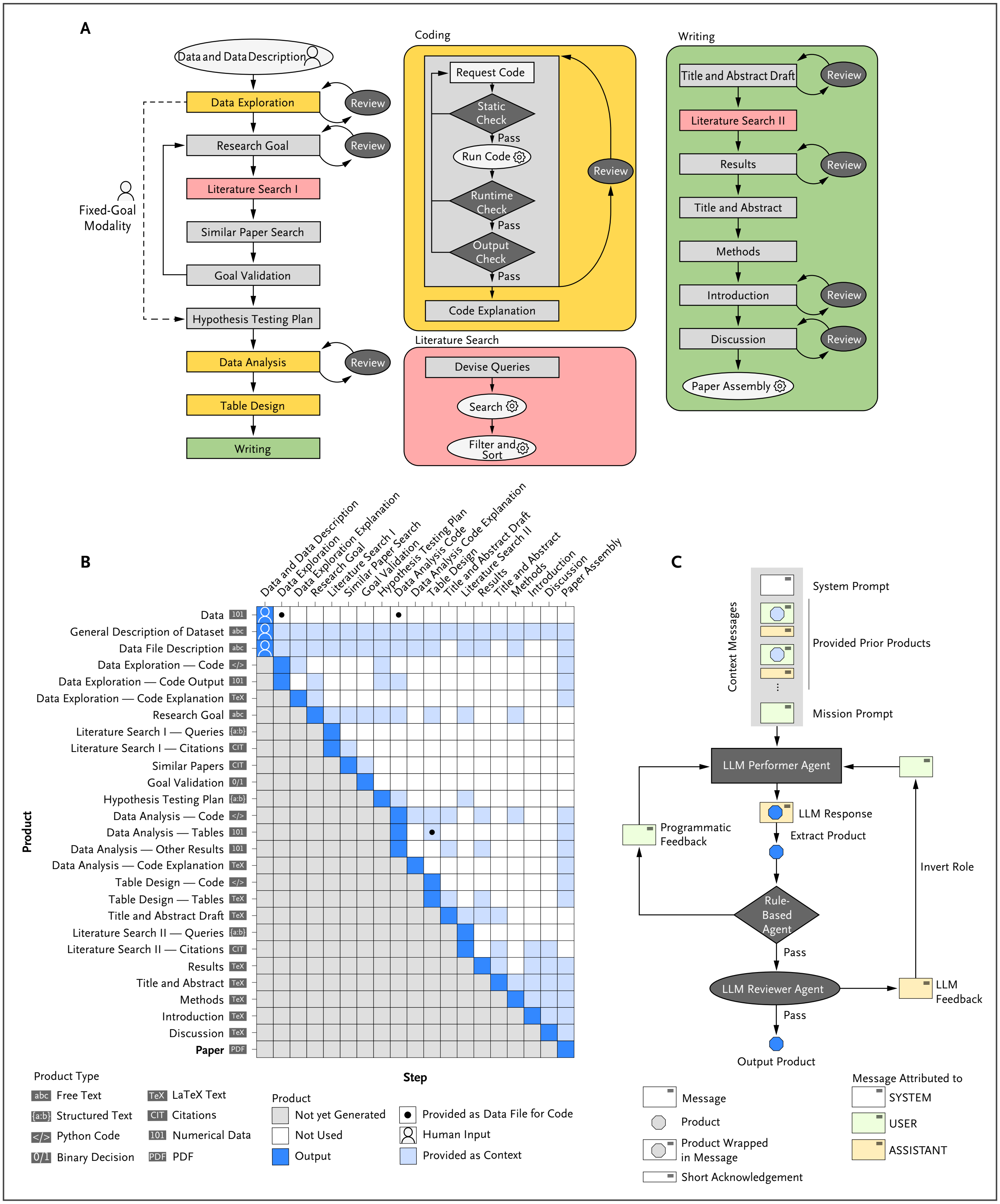


Figure 1. Data-to-Paper Orchestrates Agent Interactions and Information Flow through a Multistep Research Cycle from Annotated Data to Research Paper.

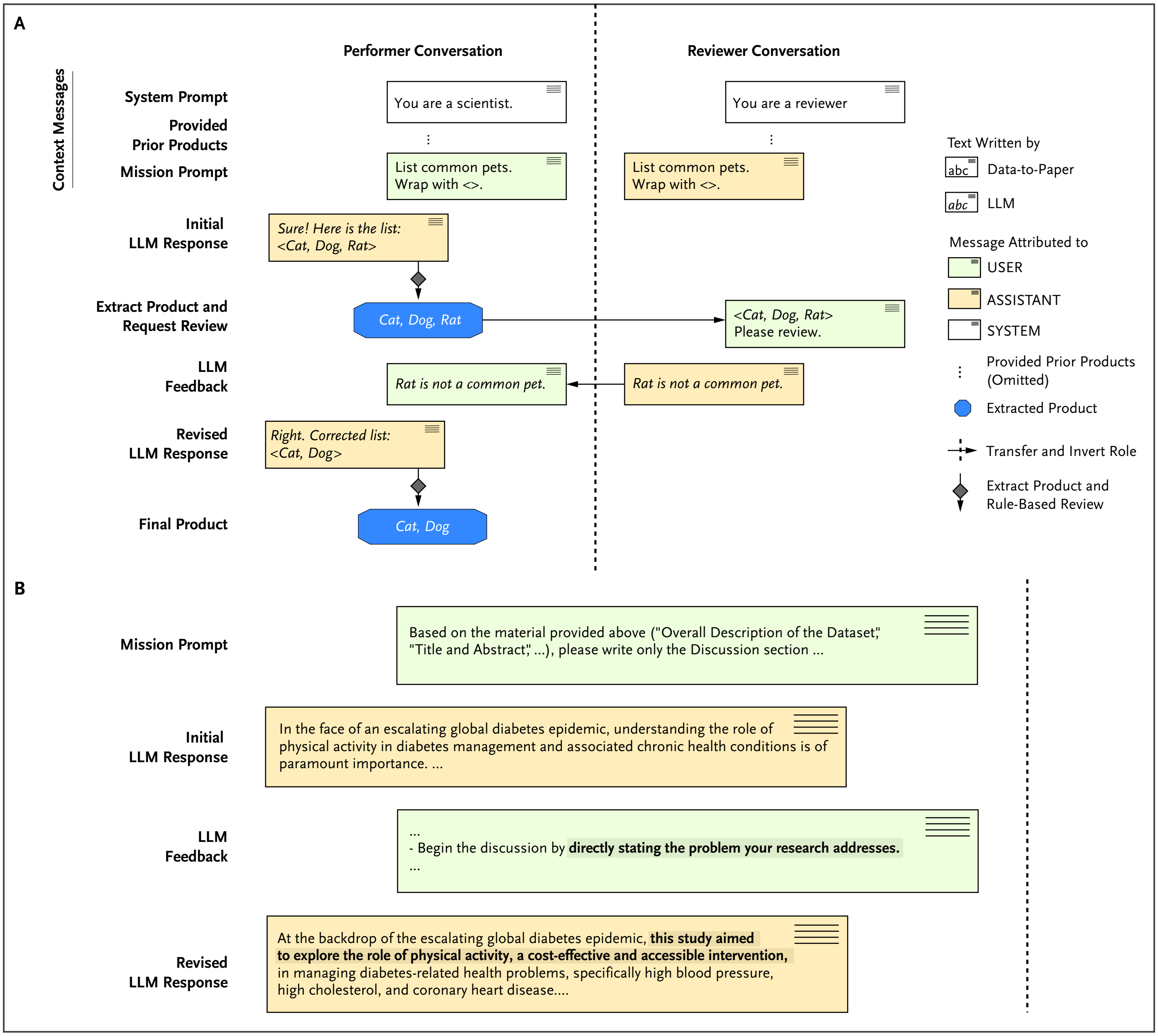
The process runs automatically through the series of steps (with human oversight and approval; Supplementary Methods), with each step creating one or more research products: free text, LaTeX text, structured text, binary decision, citations, Python code, and numerical data (Fig. 1B). In the coding steps, the LLM creates a Python code, which is then executed by data-to-paper to analyze the provided dataset and create numerical data products (such as tables for the paper; Supplementary Methods). In the literature search steps, a structured list of queries created by the LLM is used to retrieve a list of citations from an external citation database[30] (Fig. 1A; Supplementary Methods). Ultimately, these intermediate products are automatically assembled into a complete research paper (labeled with an AI-created watermark for transparency; Fig. 1B; Supplementary Methods). All steps are recorded in log files, which allow data-to-paper to fully replay the run and generate the same manuscript (Supplementary Methods).

Each research step is implemented as a distinct conversation that features agent identity specification, provision of prior research products, mission instructions, and LLM responses with iterative feedback (Fig. 1C; Supplementary Methods; Fig. S1 in Supplementary Appendix 1). First, the LLM agent is given a specific identity (Supplementary Methods; e.g., "You are a scientist who needs to write literature search queries"; performer system prompt, Table S1). Next, data-to-paper populates the conversation with a set of provided prior products, which is a list of messages that provides the LLM with a predefined subset of research products from prior steps deemed important for the focal task (Figs. 1B and S1; and provided prior products, Table S1). This rigorous control of information flow between steps minimizes possible hallucinations that can otherwise result from the mixing of relevant and irrelevant information.[31] It also allows data-to-paper to trace, verify, and chain the sources of numeric results cited by the LLM (Supplementary Methods).

Next, a step-specific mission prompt message is appended, defining the new product that the LLM is expected to create (e.g., "Please write literature-search queries…"; performer mission prompt, Table S1). Then, data-to-paper requests a response from the LLM model application programming interface,[32-34] from which it extracts the requested product (based on defined formatting; Fig. S1; Table S1; Supplementary Methods). The extracted product then undergoes a series of rule-based algorithmic checks, providing constructive feedback to the LLM upon failure (Supplementary Methods; Figs. 1C and S1). To minimize errors in the coding steps, we built a unique framework that imposes guardrails against commonly observed coding and statistical analysis errors and bad coding practices by using a series of static code checks, runtime errors, package-specific guardrails, and output verifications (Fig. 1A, coding block; Supplementary Methods).

Once the created product passes the rule-based review, it may be refined further through LLM review (inspired by other multi-agent frameworks)[9,35-40] as well as through human review (in copilot mode; steps with review ovals, Fig. 1A; Supplementary Methods). LLM review is implemented as a parallel, role-inverted conversation, effectively creating an exchange between two LLM agents

Figure 1. (*Continued*) In Panel A, starting with a human-provided dataset and its textual description (oval with human icon), data-to-paper executes a series of LLM research steps (boxes) and programmatic tools (ovals with gear icon), progressing toward the creation of a research paper. In fixed-goal modality, the research goal is human provided, and the goal-determining steps are skipped (dashed bypass arrow). Coding (yellow box), writing (green box), and literature search (red box) are modules consisting of several LLM steps and programmatic tools (Supplementary Methods in Supplementary Appendix 1). Each step creates a research product that undergoes rule-based review; an example of rule-based review is shown only for the coding step, where static, runtime (including package-specific), and output checks are performed (diamond-shaped decision points; Supplementary Methods). Some of the steps also incorporate an LLM review (review, dark ovals Supplementary Methods). In Panel B, each research step (columns) creates one or more research products (rows; output, blue), while using a provided subset of previously created products as inputs (provided as context, light blue). Products can be of different textual, structural, or numeric types (legend). Numeric products can be provided not only as conversation context, but also as data files for code (centered dot). The raw data files and their descriptions are provided as human input products (human icon). In Panel C, each research step is implemented as a distinct LLM performer conversation (Supplementary Methods), programmatically filled with context messages, including: a system prompt message defining the agent role (white message; Table S1), provided prior products, a series of USER-side messages (green) containing prior products (light blue octagons), each followed by a short LLM-surrogating ASSISTANT-side acknowledgment message (slim orange message), and a step-specific mission prompt requesting the focal product (Table S1; Figs. S1 and S4; Supplementary Methods). This prefilled conversation is passed to an LLM performer agent, which generates a response from which the requested product (blue octagon) is then extracted. This product undergoes rule-based review (dark diamond) and LLM review (dark oval), where a response from another LLM agent is cast as if it were a USER-side response (invert role; Figs. S2 and S6; Supplementary Methods). In copilot mode, human review is requested after the LLM review is completed (not shown). LLM denotes large language model; and PDF, portable document format.



Based on the material provided above ("Overall Description of the Dataset," "Title and Abstract," ...), please write only the Discussion section ...

In the face of an escalating global diabetes epidemic, understanding the role of physical activity in diabetes management and associated chronic health conditions is of paramount importance. ...

...
- Begin the discussion by **directly stating the problem your research addresses.**
...

At the backdrop of the escalating global diabetes epidemic, **this study aimed to explore the role of physical activity, a cost-effective and accessible intervention,** in managing diabetes-related health problems, specifically high blood pressure, high cholesterol, and coronary heart disease....

Figure 2. Exchange among LLM Agents Helps to Refine Research Products.

Panel A is a schematic illustration of an LLM review. The LLM review is implemented using role-inverting transfer of messages among parallel performer (left) and reviewer (right) conversations. Both conversations are initiated with an identity-defining system prompt (system prompt, white box), followed by relevant prior products (provided prior products, three dots, explained in Supplementary Methods, and Figs. 1C and S1 in Supplementary Appendix 1) and a mission prompt requesting the research product (mission prompt). Then, an initial response containing the product is returned (initial LLM response), from which the requested product is extracted and undergoes rule-based review (arrow with dark diamond; Supplementary Methods). The extracted product from the ASSISTANT-side message is then transferred to the reviewer conversation, where it is wrapped as a USER-side message (transfer and invert role, horizontal arrow). The LLM reviewer agent is replying with feedback (LLM feedback, orange box) that gets role-inverted and transferred back to the performer conversation (LLM feedback, green box). The product is refined according to the feedback (revised LLM response, orange box) and gets extracted (final product). Panel B is an example of an interaction between performer and reviewer during discussion writing (from Supplementary Run A5 in Supplementary Appendix 2). An initial draft of a discussion paragraph (initial LLM response) receives reviewer comments with suggestions for improvements (bold and highlighted text, LLM feedback), leading to textual improvements (bold and highlighted text, revised LLM response). LLM denotes large language model.

(Supplementary Methods; Figs. 1C, 2A and S2). In copilot mode, the user can provide additional review comments, resulting in further LLM iterations (Supplementary Methods). Once a product passes the rule-based review, LLM review, and, optionally, human review, the step is concluded, and data-to-paper proceeds to the next step, until all products are created, and the paper is assembled.

Although data-to-paper can work with any LLM, in the provided case studies, we used the ChatGPT model family (from ChatGPT3.5 to ChatGPT4). Using current state-of-the-art open-source LLMs, such as models from the Llama-2 model family, led to frequent mistakes that precluded completing full research cycles (Supplementary Methods; Table S2; Fig. S3). Of note, since ChatGPT models are not deterministic, each research cycle run of data-to-paper — even on the same dataset and with or without a human-provided goal — unfolded with different analyses, yielding different overall manuscripts. All technical details can be found in the Supplementary Methods.

Results

OPEN-GOAL RESEARCH ON PUBLIC DATASETS

Running in an open-goal, autopilot modality (i.e., letting LLM agents define the research goal in an iterative process), data-to-paper was given two relatively simple, publicly available datasets: the health indicators dataset,[25] an unweighted curated subset of the Centers for Disease Control and Prevention's Behavioral Risk Factor Surveillance System from 2015,[41] with 253,680 clean responses, each including 22 features related to diabetes and general health (single file of tabular data); and the social network dataset,[26] a directed graph representing Twitter, now known as X, interactions among members of the 117th U.S. congress, as well as member affiliations (two files, indicating node features and the links between them). For each of the two datasets, we ran data-to-paper for five full research cycles to create 10 distinct manuscripts (Supplementary Dataset A and B; Data Descriptions A and B; and Manuscripts A and B in Supplementary Appendix 2).

During the open-goal research cycles, which took about an hour each, data-to-paper generated and corrected hypotheses, created and debugged code, composed search queries and retrieved citations, and wrote and revised the manuscript section by section (see full conversations in Supplementary Run Files A and B; Figs. 2B and S4–S7). All manuscripts that were created followed the canonical structure of a research paper. These manuscripts included an appropriate title and abstract; a well-formulated introduction that stressed the research questions in light of relevant literature; a methods section that provided a transparent and human-traceable description of the analysis and key methodologies; several supplementary sections that provided all custom-written codes; properly formatted scientific tables; a results section that described the findings while referring to each of the tables; and a referenced discussion that summarized the results, delineated limitations, and put the findings in a broader context (Supplementary Manuscripts A1–5 and B1–5). Although similar in structure, the five papers produced for each dataset addressed different topics and raised and tested different hypotheses (Tables 1 and S3). These papers were not highly creative, yet they did define a reasonable set of hypotheses, tested them with simple straightforward statistical approaches, and ultimately created and adequately reported de novo insights from the provided data.

By manually vetting the data analysis and the text of these papers, we found that of the 10 open-goal papers (5 for each of these simple datasets), 8 reported correct analyses with no major errors, but 2 were erroneous, showing fundamental analysis or interpretation mistakes (Supplementary Manuscripts A and B; Fig. 3A). The analyses in all five health indicators papers were based on either logistic or linear regression models, and all performed adequately while accounting for a reasonable choice of confounding factors (Table S4). Moreover, interaction terms were added when needed, and the dataset was adequately restricted to reflect the tested hypotheses (e.g., restricted to a diabetic subpopulation; Table S4). For the social network dataset, papers were based on linking graph properties with node properties, as well as on creating new node properties (e.g., state representative count), and then applying linear regression, analysis of variance, or chi-square tests on either the graph nodes or edges as appropriate (Table S4; see methods sections and analysis codes in each of the created papers, Supplementary Manuscripts A1–5 and B1–5).

In all 10 papers, the generated scientific tables correctly represented the results of the analysis. Vetting the text, we observed that data-to-paper adequately interpreted the analysis results with factual statements, correctly referred to tables and cited key numeric values from the analysis, and reasonably described the research question and findings in the context of existing literature (green highlights, Supplementary Manuscripts A1–5 and B1–5; Methods). We also detected multiple imperfections, such as generic

Table 1. Examples of Topics and Findings of Papers Produced for Health Indicators Dataset (A1–3) and Social Network Dataset (B1–3).*

Paper	Topic and Findings
A1	Topic: Diabetes and physical activity Title: "Insights into the Relationship between Physical Activity and Diabetes Prevalence" Conclusion: "[...] a negative association between physical activity and diabetes prevalence, independent of confounders such as age, smoking status, and education level."
A2	Topic: Physical activity and glycemic control in diabetic population Title: "Impact of Diabetes on Physical Activity, BMI, and Demographic Factors in a Large-scale Population Study" Conclusion: "[...] distinct patterns in physical activity levels, BMI, age distribution, and sex proportions between individuals with and without diabetes."
A3	Topic: Diabetes and diet Title: "The Impact of Fruit and Vegetable Consumption on Diabetes Prevalence: Insights from a Nationwide Survey" Conclusion: "[...] significant inverse relationship between fruit and vegetable consumption and diabetes prevalence, [...]"
B1	Topic: Congress chamber and Twitter interactions Title: "Discovering Communication Patterns in the U.S. Congress through Twitter Interactions" Conclusion: "[...] uncovers a significant association between the House of Representatives and the Senate regarding Twitter interactions."
B2	Topic: Party affiliation and Twitter interactions Title: "Party Dynamics in Twitter Interactions among Members of the 117th U.S. Congress" Conclusion: "[...] a significant association between party affiliation and Twitter interactions, revealing higher levels of engagement within party lines."
B3	Topic: Home state, party affiliation and Twitter interaction Title: "Insights into Social Dynamics among U.S. Congress Members through Twitter Interactions" Conclusion: "[...] significant distinctions in Twitter interactions among different political parties, [...] unveil the influential role of represented states."

*BMI denotes body mass index; the body mass index is the weight in kilograms divided by the square of the height in meters; Twitter is now known as X.

phrasing, overstatement of novelty, and inadequate citations (yellow and orange highlights, Supplementary Manuscripts A1–5 and B1–5). Major mistakes, which affected results, were found in 2 of the 10 papers: In one of the health indicators papers, the results of a statistically sound analysis were misinterpreted due to hallucinations in the goal-specification step, leading to conclusions beyond the scope of the analysis and underlying dataset; and in one of the social network papers, an erroneous analysis was performed, resulting in unfounded statements on statistical associations between social interactions and party affiliations (red highlights, Supplementary Manuscripts A2 and B2, respectively).

To test data-to-paper with more complex and challenging data, we chose a dataset of SARS-CoV-2 infection events in healthcare workers with different vaccination statuses. The dataset contained uneven and unbalanced participant time series (infection dataset,[27] two files containing a number of symptoms and viral variants and the time intervals with medical events). Running data-to-paper on this dataset in autopilot, we found that it made major analysis mistakes related to data handling, whereby it incorrectly merged the two files or aggregated the time intervals, and thereby causing a data distortion (Fig. 3A; red highlights in data analysis code, Supplementary Manuscripts C1–4; Run Files C1–4). Yet, with human copiloting, data-to-paper was able to overcome these data handling issues and generate correct analyses and sound papers, recapitulating aspects of prior analyses of the data[27] (see example Supplementary Manuscript Ch and Run File Ch, where ~15 short human review comments were provided, related to both analysis and interpretation). Data-to-paper can thus autonomously tackle simple datasets in autopilot mode, but it requires human copiloting for correct analysis of complex datasets.

ESTIMATING RELIABILITY IN REPRODUCING PEER-REVIEWED RESULTS

To more systematically assess the error rate in autopilot and copilot modes, we applied data-to-paper to two case studies for which we had benchmarks of published, peer-reviewed results. We specifically wanted to check two critical aspects regarding the reliability of analysis and interpretation: the proper reporting of both positive and negative findings (challenge 1), and the performance for tasks with multiple different steps with tunable breadth (challenge 2). To test the capacity of data-to-paper to address these two challenges, we chose two peer-reviewed studies: a study by Saint-Fleur et al.,[28] which adequately reports both positive and negative findings related to the association of a policy change in a neonatal intensive care unit with treatment choice and treatment outcome (challenge 1); and a study by

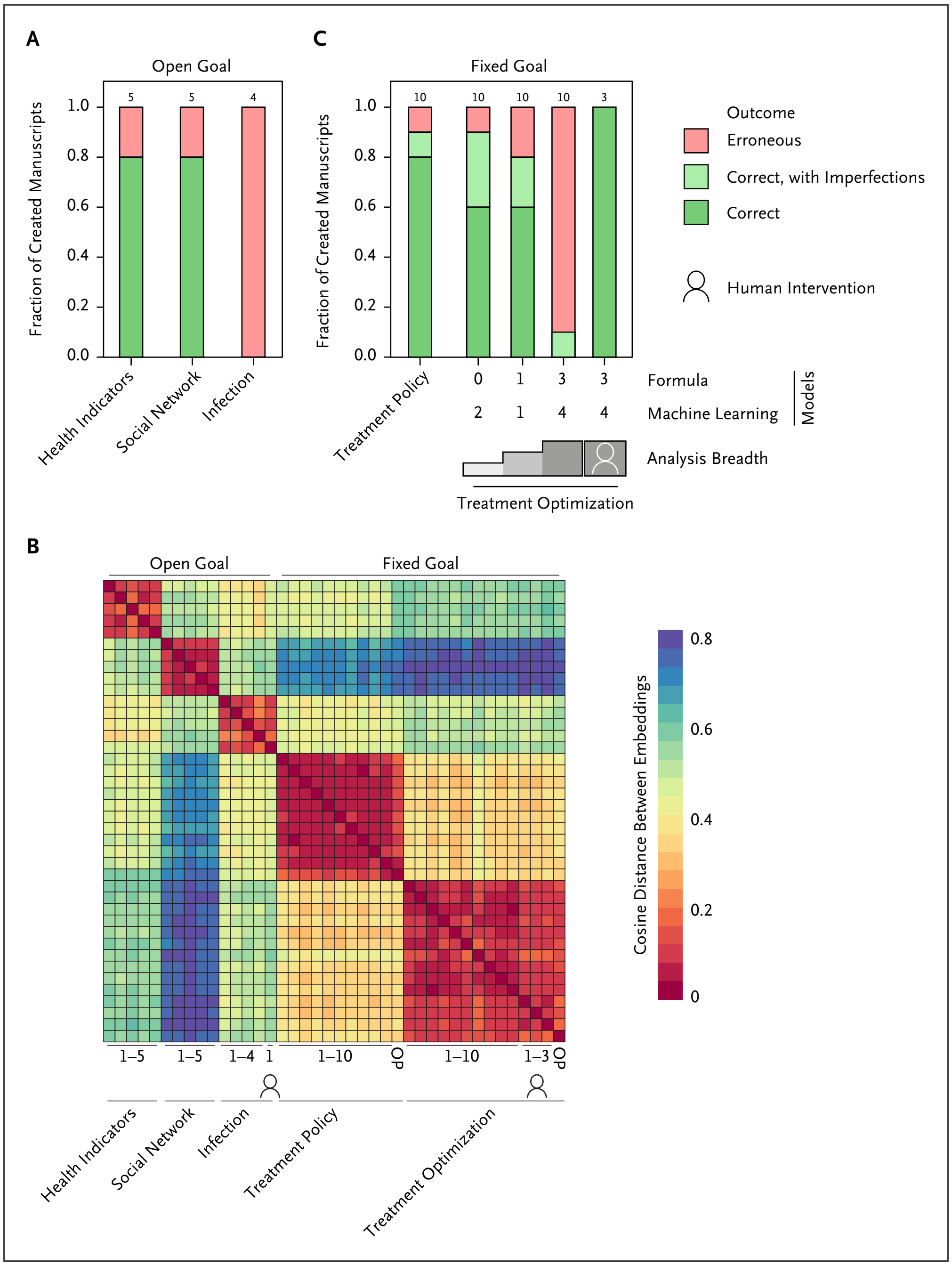


Figure 3. Data-to-Paper Is Able to Autonomously Create Correct Papers for Simple Research Goals and Datasets While Human Copiloting Is Required to Ensure Accuracy in More Complex Settings.

In Panel A, for manuscripts generated in open-goal modality, a bar plot showing the fraction of correct (green), correct with imperfections (light green), and erroneous (red) papers that were generated for each of the three datasets. The numbers above the bar indicate the number of papers created by data-to-paper for each set. In Panel B, embedding vectors of the title and abstract of each paper were generated using SPECTER.[42] Heatmap showing pairwise cosine distance between embedding vectors of the 5 health indicators papers, 5 social network papers, 5 infection papers, 10 treatment policy papers, 13 treatment optimization papers, and the 2 original papers.[28,29] Human icon indicates human intervention. In Panel C, similar to Panel A, for manuscripts generated in fixed-goal modality for the treatment optimization dataset; three sets of papers were created, corresponding to research goals of varying complexity (Supplementary Methods; Data Descriptions Ea, Eb and Ec; Fig. S8). For the most complex goal, data-to-paper was also run in copilot mode (human icon; Supplementary Manuscripts Eh1–3). For the annotation of manuscripts for imperfections and errors see highlighted text in Supplementary Manuscripts A1–5, B1–5, C1–4, D1–10, Ea1–10, Eb1–10 and Ec1–10. OP denotes original papers.

Shim et al.,[29] which builds several machine learning models for predicting optimal intubation depth in pediatric patients and compares their prediction accuracy with formula-based models; the task requires multiple analysis steps, whose breadths can be gradually tuned by altering the number of models to compare (challenge 2). Both studies provided well-annotated datasets, and both were published after the knowledge cutoff date of the ChatGPT models that we used to produce them (September 2021; Supplementary Methods), such that the findings of the studies were not part of the LLM's training data. Therefore, its conclusions could not be simple retrievals from the training data. When data-to-paper was run in open-goal modality, it typically chose research goals that were similar to or simpler than those in the corresponding original studies, which led to correct but often less interesting manuscripts (see Supplementary Manuscripts Do and Eo as examples).

To ensure a direct comparison with the original studies, we next ran data-to-paper with these two datasets in fixed-goal modality and provided it with the corresponding research goals of the original publication and ran it for 10 independent research cycles (Supplementary Data Descriptions D and E; Datasets D and E; Tables S5 and S6; Manuscripts D1–10 and Ea1–10; and Run Files D and E). For each case study, the created papers were similar to each other in terms of content, terminology, and structure. Indeed, quantifying content similarity by the pairwise cosine distance between the vector embeddings of the title and abstract of all created manuscripts[42] showed tight and distinct clusters corresponding to the five case studies (Fig. 3). These fixed-goal papers were also similar to their respective original studies[28,29] in content, terminology, and vector embeddings (Fig. 3B).

We manually vetted the analysis and reported results of the manuscripts created for each of the two study-reproducing challenges. For challenge 1, we found that all papers correctly reproduced the analysis, and eight of them reached the overall correct conclusions and adequately reported both the negative and positive results. All of these manuscripts used adequate statistical methodologies, either matching the methods used in the original study[28] or providing valid alternatives (Table S5; Supplementary Manuscripts and Run Files D). However, in two of the papers we identified interpretation errors, which also affected the overall conclusions in one of the papers (Fig. 3C; Supplementary Manuscripts D1 and D2, red and orange highlights; Tables S5 and S6). For challenge 2, we found that the rate of error varied critically with the breadth of the analysis. Although data-to-paper frequently failed (90% error rate) when presented with the original, broad research goal, it was able to correctly perform this multistep model development research for almost identical research goals, except for requesting to develop and compare fewer models (10 to 20% error rate; Figs. 3C and S8).

We noted that in all cases, process reliability depends on the formulation of the research goal and on a well-structured, concise description of the dataset; less detailed and explicit formulations can increase analysis errors (Fig. S8; Supplementary Manuscripts Eai1–10, Ebi1–10 and Eci1–10 and Data Descriptions Eai, Ebi and Eci; and Run Files E). Finally, allowing human copiloting (Supplementary Methods) with a few brief review comments per run, typically in the code writing step, enabled data-to-paper to consistently create accurate papers, even for those with complex goals (Fig. 3; Supplementary Manuscripts Eh1–3). Altogether, these case studies provide an assessment of data-to-paper's analysis and interpretation reliability, showing that for simple research goals, it can autonomously create reliable manuscripts in 80 to 90% of cases, but that for more complex goals, human copiloting is critical to ensure reliability.

Finally, noting the effort and necessity of manually vetting and verifying created manuscripts, we harnessed and enhanced data-to-paper's step-to-step information tracing to actively chain results, methodology, and data in created manuscripts through algorithmically verified hyperlinks (Supplementary Methods, see specific data-chained Supplementary manuscripts, e.g., Ch; note that previously generated manuscripts are not data-chained). This approach creates manuscripts in which all cited numeric values are recursively linked to the specific lines of code where they are created. In particular, each numerical value cited in the manuscript is linked to all intermediate upstream products, such as a notes appendix that provides the formula and explanation for values that were transformed in the text; the specific table from which values used in these formulae originated; the corresponding output file of the code from which the table was created; and to the specific part of code that produced this output file (see hyperlinks in data-chained Supplementary Manuscripts, e.g., Ch; Supplementary Video in Supplementary Appendix 2). Such data-chained manuscripts facilitate systematic vetting of papers and set a new standard of traceability and verifiability for the coming era of AI-powered research.

Discussion

Inspired by key features of human research, we combined ideas from prompt automation, tool augmentation, and multiagent interaction approaches[9,12,35] to guide multiple LLM agents through a full research path from annotated data to well-structured, transparent, human-verifiable papers. Tracing information through the different research steps allows data-to-paper to create data-chained manuscripts, in which results, methodologies, and data are programmatically linked. Although the novelty of current AI-driven research falls well behind that of high-end contemporary science, our platform demonstrates the de novo creation of new insights from provided data, thereby mimicking a key aspect of human research and taking science automation well beyond what is possible with algorithmic data exploration.[43] Moreover, the process demonstrates versatility with respect to data types and research domains, and is able to perform different types of scientific research, such as association studies, network analysis, and the development and testing of machine learning models.

However, when run fully autonomously, the process is not error-free; despite attempts to minimize errors with multiple guardrails, algorithmic checks, review cycles, and tight control of information flow, the notorious problem of LLM hallucinations[31] and the inability to properly tackle complex datasets led to fundamental errors in about 10 to 20% of created papers for simple analysis tasks and simple datasets, as well as to consistent failure in more complex tasks or with more complex datasets. Failures mainly included data handling or statistical errors, although an instance of a nonsensical hallucination statement was also observed. With human copiloting, several short review comments were sufficient to overcome these failure modes, possibly delineating best practices for real-world applications of platforms like data-to-paper. Of note, the effectiveness of human copiloting strongly depends on human expertise and highlights the importance of using domain experts in AI-assisted research to ensure accuracy and overall quality.

Our current implementation has several constraints: it is limited to textual and table outputs, is unable to pursue follow-up questions, and is restricted to hypothesis-testing research on preexisting data. Despite these current limitations, which are expected to be overcome with further developments in LLMs and refinements of our approach, the ability of LLMs to carry out scientific research presents important opportunities. Indeed, such AI research approaches could dramatically accelerate the pace of scientific research, especially in fields such as epidemiology and biomedicine, where new data are rapidly generated, often at a pace exceeding the capacity of current research processes. Integrating data-to-paper or similar emerging AI-driven science platforms[44] into the routine workflow of medical or public health agencies, such as surveillance bodies, health insurance providers, and hospitals, could allow tapping into an underexploited source of data. Potentially, our platform might be integrated with concurrent efforts into the development of LLM-driven hypothesis generation or paper review.[45-47]

Importantly, the creation of data-chained manuscripts demonstrates that AI-driven science does not necessarily jeopardize research traceability and verifiability; on the contrary, it can help enhance them even beyond the standard of human-driven research. However, there are also risks associated with this development, such as the dishonest use of such systems, for example, in the context of *P*-hacking[48] or of overloading the publication system with medium-level and generic manuscripts that address insignificant problems.[49-51]

Our approach implements specific features to mitigate some of these risks, in line with emerging guidelines on AI in science,[23] including a transparent oversight process that allows human copiloting; unbiased reporting of either positive or negative results; the creation of transparent, AI-marked, data-chained, and human-verifiable papers; and a complete and fully replayable recording of each run. Given the relatively limited novelty and the potential for errors in fully autonomous AI-driven research, as well as the need for ethical judgments and accountability,[23] we anticipate and urge that such AI-driven approaches will be used as scientist copilots, helping scientists in the more straightforward tasks, thereby allowing them to focus their minds and creativity on higher-level concepts.

Disclosures

Author disclosures and other supplementary materials are available at ai.nejm.org.

We thank Eric Lander, Yoel Fink, and Michael Elowitz for insightful discussions; Ayelet Baram-Tsabari and Yael Rozenblum for discussions and providing data for initial tests; Ofer Sapir for help in organizing data-to-paper repository; and all laboratory members for their helpful comments. Dr. Hafner was supported in part at Technion–Israel Institute of Technology by Aly Kaufman and Swiss National Science Foundation Postdoc.Mobility Fellowships.

The data that support the findings of this study are available in the article and its Supplementary Appendix 1 and 2.

Code is available at https://github.com/Technion-Kishony-lab/data-to-paper

Author Affiliations

[1] Faculty of Data and Decision Sciences, Technion–Israel Institute of Technology, Haifa, Israel

[2] Faculty of Biology, Technion–Israel Institute of Technology, Haifa, Israel

[3] Epsio, Tel Aviv, Israel

[4] Faculty of Computer Science, Technion–Israel Institute of Technology, Haifa, Israel

[5] Faculty of Biomedical Engineering, Technion–Israel Institute of Technology, Haifa, Israel

References

1. Xi Z, Chen W, Guo X, et al. The rise and potential of large language model based agents: a survey. September 14, 2023 (http://arXiv.org/abs/2309.07864). Preprint.

2. Liu Y, Han T, Ma S, et al. Summary of ChatGPT-related research and perspective towards the future of large language models. Meta-Radiology 2023;1:100017. DOI: 10.1016/j.metrad.2023.100017.

3. Wang L, Ma C, Feng X, et al. A survey on large language model based autonomous agents. August 22, 2023 (http://arXiv.org/abs/2308.11432). Preprint.

4. Li Y, Choi D, Chung J, et al. Competition-level code generation with AlphaCode. Science 2022;378:1092-1097. DOI: 10.1126/science.abq1158.

5. Spitale G, Biller-Andorno N, Germani F. AI model GPT-3 (dis)informs us better than humans. Sci Adv 2023;9:eadh1850. DOI: 10.1126/sciadv.adh1850.

6. He-Yueya J, Poesia G, Wang RE, Goodman ND. Solving math word problems by combining language models with symbolic solvers. April 26, 2023 (http://arXiv.org/abs/2304.09102). Preprint.

7. Romera-Paredes B, Barekatain M, Novikov A, et al. Mathematical discoveries from program search with large language models. Nature 2024;625:468-475. DOI: 10.1038/s41586-023-06924-6.

8. Trinh TH, Wu Y, Le QV, Luong T. Solving olympiad geometry without human demonstrations. Nature 2024;625:476-482. DOI: 10.1038/s41586-023-06747-5.

9. Qian C, Cong X, Liu W, et al. Communicative agents for software development. July 16, 2023 (http://arXiv.org/abs/2307.07924). Preprint.

10. Dong Y, Jiang X, Jin Z, Li G. Self-collaboration code generation via ChatGPT. April 15, 2023 (http://arXiv.org/abs/2304.07590). Preprint.

11. Zhou W, Jiang YE, Cui P, Kannan A. RecurrentGPT: interactive generation of (arbitrarily) long text. May 22, 2023 (http://arXiv.org/abs/2305.13304). Preprint.

12. Nair V, Schumacher E, Tso G, Kannan A. DERA: enhancing large language model completions with dialog-enabled resolving agents. March 30, 2023 (http://arXiv.org/abs/2303.17071). Preprint.

13. Tu T, Palepu A, Schaekermann M, et al. Towards conversational diagnostic AI. January 11, 2024 (http://arXiv.org/abs/2401.05654). Preprint.

14. Boiko DA, MacKnight R, Kline B, Gomes G. Autonomous chemical research with large language models. Nature 2023;624:570-578. DOI: 10.1038/s41586-023-06792-0.

15. Liang W, Zhang Y, Cao H, et al. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. NEJM AI 2024;1(8). DOI: 10.1056/AIoa2400196.

16. Birhane A, Kasirzadeh A, Leslie D, Wachter S. Science in the age of large language models. Nat Rev Phys 2023;5:277-280. DOI: 10.1038/s42254-023-00581-4.

17. Conroy G. How ChatGPT and other AI tools could disrupt scientific publishing. Nature 2023;622:234-236. DOI: 10.1038/d41586-023-03144-w.

18. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. Nature 2023;614:214-216. DOI: 10.1038/d41586-023-00340-6.

19. Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove. Nature 2023;613:620-621. DOI: 10.1038/d41586-023-00107-z.

20. Hutson M. Could AI help you to write your next paper? Nature 2022;611:192-193. DOI: 10.1038/d41586-022-03479-w.

21. Berdejo-Espinola V, Amano T. AI tools can improve equity in science. Science 2023;379:991. DOI: 10.1126/science.adg9714.

22. Messeri L, Crockett MJ. Artificial intelligence and illusions of understanding in scientific research. Nature 2024;627:49-58. DOI: 10.1038/s41586-024-07146-0.

23. Bockting CL, van Dis EAM, van Rooij R, Zuidema W, Bollen J. Living guidelines for generative AI — why scientists must oversee its use. Nature 2023;622:693-696. DOI: 10.1038/d41586-023-03266-1.

24. Wilson EB. An introduction to scientific research. McGraw-Hill Book Company, 1952.

25. Teboul A. Diabetes health indicators dataset. 2021 (https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset).

26. Fink CG, Omodt N, Zinnecker S, Sprint G. A congressional twitter network dataset quantifying pairwise probability of influence. Data Brief 2023;50:109521. DOI: 10.1016/j.dib.2023.109521.

27. Babouee Flury B, Güsewell S, Egger T, et al. Risk and symptoms of COVID-19 in health professionals according to baseline immune status and booster vaccination during the Delta and Omicron waves in Switzerland-a multicentre cohort study. PLoS Med 2022;19:e1004125. DOI: 10.1371/journal.pmed.1004125.

28. Saint-Fleur AL, Alcalá HE, Sridhar S. Outcomes of neonates born through meconium-stained amniotic fluid pre and post 2015 NRP guideline implementation. PLoS One 2023;18:e0289945. DOI: 10.1371/journal.pone.0289945.

29. Shim J-G, Ryu K-H, Lee SH, Cho E-A, Lee S, Ahn JH. Machine learning model for predicting the optimal depth of tracheal tube insertion in pediatric patients: a retrospective cohort study. PLoS One 2021;16:e0257069. DOI: 10.1371/journal.pone.0257069.

30. Kinney R, Anastasiades C, Authur R, et al. The semantic scholar open data platform. 2023 (http://arXiv.org/abs/2301.10140). Preprint.

31. Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. February 8, 2022 (http://arXiv.org/abs/2202.03629). Preprint.

32. OpenAI Platform (https://platform.openai.com/docs/api-reference).

33. Touvron H, Martin L, Stone K, et al. Llama 2: open foundation and fine-tuned chat models. July 18, 2023 (http://arXiv.org/abs/2307.09288). Preprint.

34. Rozière B, Gehring J, Gloeckle F, et al. Code Llama: open foundation models for code. August 24, 2023 (http://arXiv.org/abs/2308.12950). Preprint.

35. Wu Q, Bansal G, Zhang J, et al. AutoGen: enabling next-gen LLM applications via multi-agent conversation. August 16, 2023 (http://arXiv.org/abs/2308.08155). Preprint.

36. Hong S, Zhuge M, Chen J, et al. MetaGPT: meta programming for a multi-agent collaborative framework. August 1, 2023 (http://arXiv.org/abs/2308.00352). Preprint.

37. Madaan A, Tandon N, Gupta P, et al. Self-refine: iterative refinement with self-feedback. March 30, 2023 (http://arXiv.org/abs/2303.17651). Preprint.

38. Harrison C. LangChain. 2022 (https://github.com/langchain-ai/langchain).

39. Gravitas S. AutoGPT. 2023 (https://github.com/Significant-Gravitas/AutoGPT).

40. Peng B, Galley M, He P, et al. Check your facts and try again: improving large language models with external knowledge and automated feedback. February 24, 2023 (http://arXiv.org/abs/2302.12813). Preprint.

41. Rolle-Lake L, Robbins E. Behavioral risk factor surveillance system. In: StatPearls. Treasure Island, FL: StatPearls Publishing, 2023 (https://www.ncbi.nlm.nih.gov/pubmed/31971707).

42. Cohan A, Feldman S, Beltagy I, Downey D, Weld DS. SPECTER: document-level representation learning using citation-informed transformers. April 15, 2020 (http://arXiv.org/abs/2004.07180). Preprint.

43. Steinruecken C, Smith E, Janz D, Lloyd J, Ghahramani Z. The automatic statistician. Automated machine learning: methods, systems, challenges. Springer International Publishing, 2019, pp. 161-173. DOI: 10.1007/978-3-030-05318-5_9.

44. Majumder BP, Surana H, Agarwal D, et al. Data-driven discovery with large generative models. February 21, 2024 (http://arXiv.org/abs/2402.13610). Preprint.

45. Abdel-Rehim A, Zenil H, Orhobor O, et al. Scientific hypothesis generation by a large language model: laboratory validation in breast cancer treatment. May 20, 2024 (http://arXiv.org/abs/2405.12258). Preprint.

46. Wang Q, Downey D, Ji H, Hope T. SciMON: scientific inspiration machines optimized for novelty. May 23, 2023 (http://arXiv.org/abs/2305.14259). Preprint.

47. Lu C, Lu C, Lange RT, Foerster J, Clune J, Ha D. The AI scientist: towards fully automated open-ended scientific discovery. August 12, 2024 (http://arXiv.org/abs/2408.06292). Preprint.

48. Altman N, Krzywinski M. P values and the search for significance. Nat Methods 2016;14:3-4. DOI: 10.1038/nmeth.4120.

49. Van Noorden R. Hundreds of gibberish papers still lurk in the scientific literature. Nature 2021;594:160-161. DOI: 10.1038/d41586-021-01436-7.

50. Cabanac G, Labbé C. Prevalence of nonsensical algorithmically generated papers in the scientific literature. J Assoc Inf Sci Technol 2021;72:1461-1476. DOI: 10.1002/asi.24495.

51. Liverpool L. AI intensifies fight against “paper mills” that churn out fake research. Nature 2023;618:222-223. DOI: 10.1038/d41586-023-01780-w.