ACS NANO

www.acsnano.org

ARTICLE

A Generative Artificial Intelligence Copilot for Biomedical Nanoengineering

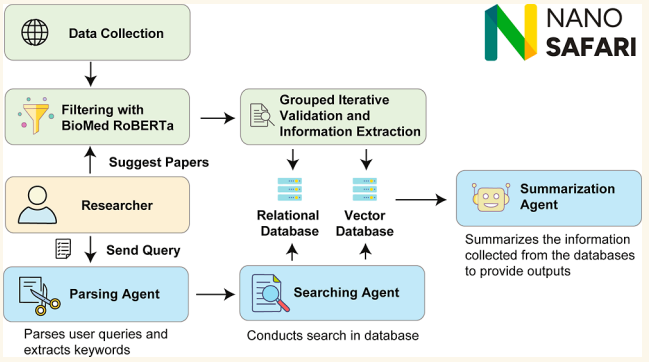
Yifan Wang,[III] Haitao Song,[III] Yue Teng,[III] Guan Huang,[III] Jingzhe Qian,[III] Hongyu Wang,[III] Shiyan Dong, JongHoon Ha, Yifan Ma, Mengyu Chang, Seong Dong Jeong, Weiye Deng, Benjamin R. Schrank, Adam Grippin, Annette Wu, Jared L. Edwards, Yixiang Zhang, Yuanyuan Lin, Wilson Poon, Stefan Wilhelm, Ye Bi, Lesheng Teng, Zikai Wang,* Betty Y. S. Kim,* and Wen Jiang*

Cite This: https://doi.org/10.1021/acsnano.5c03454

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: **The recent success of large language models (LLMs) in performing natural language processing tasks has increased interest in applying generative artificial intelligence (AI) to scientific research. However, a common problem of LLMs is their tendency to produce inaccurate and sometimes "hallucinated" outputs. Here, we established a generative AI tool, NanoSafari, to automatically extract knowledge from the biomedical nanoscience literature and address scientific queries. We developed the Grouped Iterative Validation based Information Extraction (GIVE) method to extract contextual information on nanoparticle characteristics from >20,000 published articles and established a database that was incorporated into the generative LLM to provide accurate nanomaterial design parameters. Blinded evaluation by biomedical nanoscientists showed that NanoSafari outperformed the baseline model in providing more reliable parameters for nanomaterial design tasks, as further validated by bench experiments. Together, these findings demonstrate the utility of AI-based methods for automated learning from "real-world" published work to provide accurate and reliable scientific references for biomaterial and bioengineering applications.**



KEYWORDS: *artificial intelligence, large language models, nanomedicine, nanoparticle, drug delivery*

Nanomaterials provide a powerful and highly versatile platform for drug delivery.[1] Modification of the biological, chemical, and physical characteristics of nanoparticles changes their interactions with biological systems,[2,3] and features such as material, size, and charge can affect their biodistribution.[4,5] Additional surface modifications, such as PEGylating or targeting peptides, further improve the targeted delivery of nanoparticles.[6] Various drugs and biomolecules can also be loaded into the nanoparticles. For example, several cancer chemotherapy drugs in nanoparticle formulations have been demonstrated to have superior solubility, pharmacokinetics, and treatment efficacy.[7,8] Gene therapies involving mRNA delivery by nanoparticles also show great potential in many applications, with a recent example being the mRNA-based COVID-19 vaccine, which was delivered with lipid nanoparticles.[9] Proper design of nanoparticles to achieve various goals is the central task of nanomedicine drug delivery.[10] Advances in knowledge regarding nanoparticle delivery and material science as well as the development of more sophisticated tools can help in designing novel nanomedicine with improved efficacy and safety.

Among the emerging methods to achieve precise and efficient nanomedicine design, the integration of artificial intelligence (AI) is rapidly transforming this landscape. AI and machine learning (ML) have been among the fastest evolving technologies during the past few years. AI has led to several astonishing achievements in biomedical fields, including protein structure prediction,[11] multiomic data analysis,[12] image analysis,[13] and de novo protein design.[14] Many exciting applications of AI in nanomedicine have been achieved, including analyses of images of blood vessels and nanoparticles to characterize nanoparticle delivery.[13] AI can also be trained

Received: February 25, 2025
Revised: May 4, 2025
Accepted: May 5, 2025

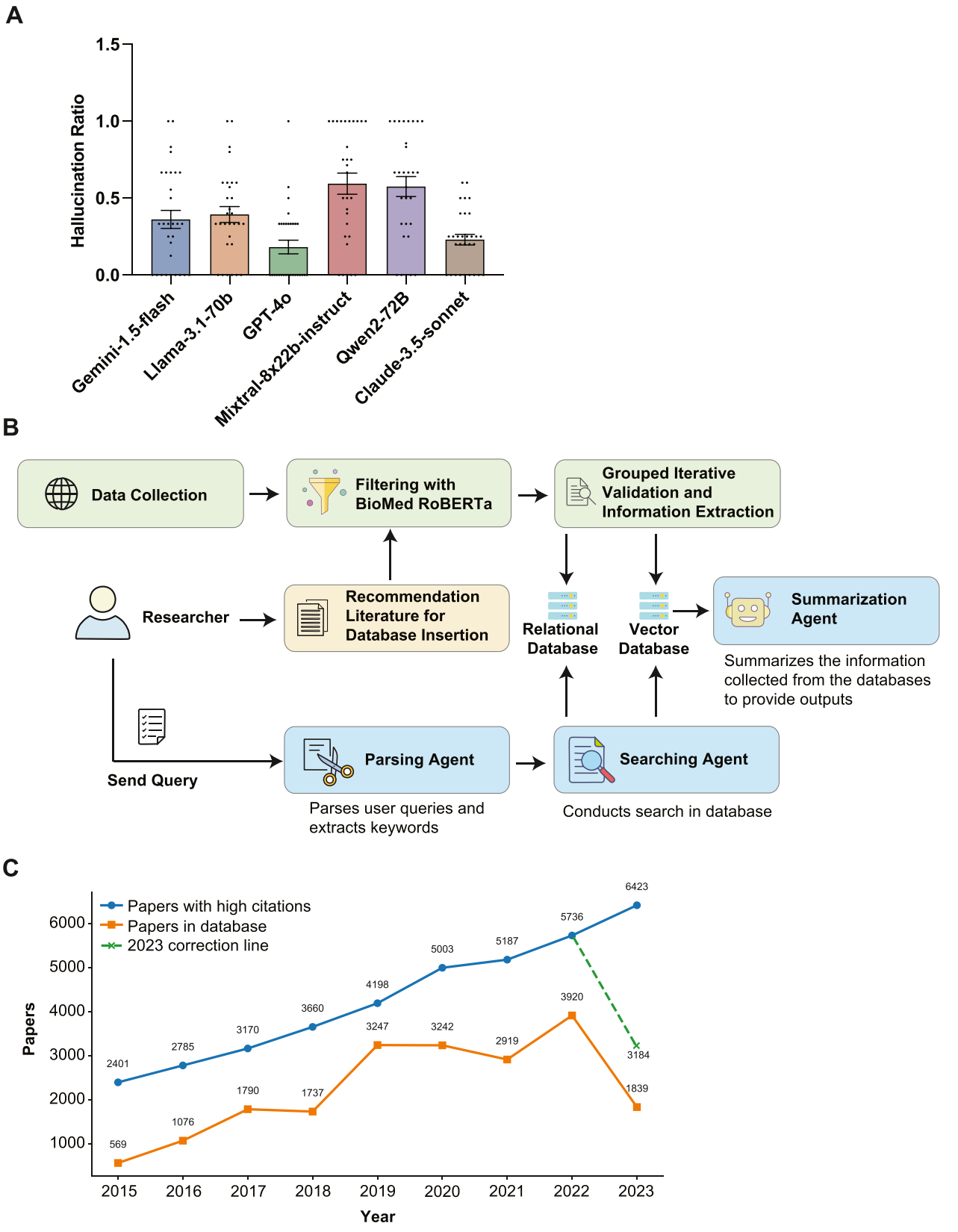


Figure 1. A novel application to increase the reliability of LLMs. (A) "Hallucinations" generated by different LLMs, shown as the invalid reference ratio of answers (n = 30). Error bars are means ± s.e. (B) Overview of NanoSafari, a two-stage system consisting of information collection and question answering. For information collection, NanoSafari can extract legitimate papers online by using external tools or accept users' paper submissions. Once papers are gathered, NanoSafari uses a fine-tuned BioMed RoBERTa model to validate whether the uploaded paper meets the requirements for the database. Subsequently, predefined attributes would be extracted by using an iterative validation extraction framework and then written into the database. For question answering, when a user submits a query, NanoSafari processes it by breaking the query down into subqueries. The system then targets specific attributes that align with the database's content in

https://doi.org/10.1021/acsnano.5c03454
ACS Nano XXXX, XXX, XXX−XXX

Figure 1. continued

the query. Using both lexical and semantic retrieval techniques, NanoSafari identifies papers that are most relevant to the query. Finally, an LLM compiles and summarizes the retrieved information, providing users with concise and relevant answers. (C) Trends in publications on nanoparticle drug delivery research from 2015 to 2023. The blue line indicates numbers of scholarly articles each year that fall within the top 50% of average annual citations for nanoparticle drug delivery; the orange line depicts the cumulative number of articles collected in our database; the green line indicates the number of articles from 2023, adjusted according to the selection criteria from previous years.

to generate a model to predict the delivery of nanomedicine to tumors.[15,16] In this context, the rise of AI technologies provides an exciting opportunity to overcome existing challenges in nanomedicine design.[17] Several platforms based on deep learning have been developed for nanoparticle design.[18,19] In addition to these exciting AI applications, another major AI tool, large language models (LLMs), is also beginning to garner increased interest in biomedical research. LLMs are AI models that have been trained on massive amounts of data, enabling them to understand and generate human-like text. LLMs have raised significant interest and attention and are extensively used in many fields. LLMs such as the Generative Pre-trained Transformer (GPT) leverage deep neural networks, particularly the Transformer architecture,[20] to process and generate text with high contextual understanding. The interactive features of LLMs allow users to query the AI and receive output with natural language. LLMs are easier to use than other AI tools and can provide more timely responses to queries, such as addressing clinical questions.[21,22] However, the current LLM models have significant drawbacks, including "hallucinations" where the model generates information not grounded in factual data, thus posing critical risks that erode the credibility of research.[23,24] Recent efforts to improve the reliability of LLMs for medical or research use include training LLMs with large medical and clinical datasets for specialized applications[25] or to fine-tune existing models.[26]

Currently, LLMs such as ChatGPT hold great potential to assist nanomedicine design by providing generic suggestions to assist researchers, such as summarizing general knowledge of a design feature, or providing suggestions to optimize nanoparticle formulation. Many researchers have started to use LLMs, in replacement of or in addition to traditional searching tools like Google or PubMed, to find solutions for their problems. However, in part due to the inherent limitations of LLMs, unreliable and fabricated outputs remain prevalent. To improve the current LLMs and enable them to serve as a "copilot" in nanomedicine development, we introduce a new LLM-based nanoparticle design platform, NanoSafari. We optimized the LLM method to create a pipeline to extract key information on nanoparticle design from more than 20,000 peer-reviewed articles on nanomedicine drug delivery. The extracted information formed a database that was used for the generation of information by LLM models to provide accurate information. Users can obtain suggestions on nanoparticle design and answers about nanomedicine drug delivery questions via NanoSafari. Notably, NanoSafari provides 100% valid and relevant references in line with the outputs to assist in nanomedicine research.

RESULTS

Fake References Are a Major Problem with LLMs. With the increasing use of LLMs such as ChatGPT in biomedical research, we sought to examine the reliability of the answers generated by various LLMs. We asked some popular LLMs 30 questions related to nanomedicine and nanoparticle drug delivery and examined the validity of the references provided in the answers by manually checking whether the cited references exist. Consistent with previous reports regarding 'hallucinations' of LLMs,[27] we found that approximately 20% of the references generated by GPT-4o-mini were fabricated, with other models performing even worse (Figure 1a). Based on these findings, we decided to use GPT-4o-mini for subsequent experiments due to its lower usage cost and faster response time, offering both temporal and financial savings while maintaining a comparable architecture to GPT-4o. To overcome this limitation and make LLMs more helpful for designing nanomedicine, we sought to improve the validity of referencing by creating a basis of solid scientific data for the LLM to cite from by gathering peer-reviewed articles from the scientific literature and extracting key information from those articles (Figure 1b). To do so, we searched Clarivate Analytics' Web of Science platform with keywords (Table S1) for significant publications in the field over the past decade. We then selected the top 50% of papers with the highest average annual citations each year. To minimize the limitations associated with using average numbers of citations for evaluating recent work, we selected 2023 publications with citation counts above the mean of the previous years (Figure 1c). Papers were downloaded in Portable Document Format (PDF) manually via institutional subscription or directly from the publisher if the paper was available via open access. Given the variations in availability of various journal subscriptions, we finally obtained 20,338 highly cited papers published in 2015−2023 (Figure 1c).

Extraction of Key Information from Published Papers. Next, we defined the following parameters as essential and key information to be extracted: nanoparticle type, target site, disease, payload, animal model, surface modification, administration route, particle size, and particle charge, plus some supporting information (Table S2). To automatically extract information on nanoparticle design from the papers, we developed a method called Grouped Iterative Validation based Information Extraction (GIVE) (Figure 2a). GIVE was developed in two stages: First, we introduced a grouped attribute extraction stage, in which related attributes were categorized into broader groups based on their interdependencies (Table S3). For example, the nanoparticle characteristics, including type, size, and charge, were extracted in the same group. The animal information, such as species, age, and sex, were extracted together in another group. This grouping strategy enabled the LLM to focus on relevant paragraphs for each group, leading to more specific and detailed extractions. When extracting a specific attribute, GIVE first identifies the paragraphs related to its group and then extracts the attribute according to the prompt, while filtering out irrelevant information. The extracted answers were stored for subsequent verification. Then, we implemented an iterative validation and extraction process, drawing on the concept of a self-reflective agent.[28] The same attribute is extracted multiple times, and the verification function checks whether the newly obtained

https://doi.org/10.1021/acsnano.5c03454
ACS Nano XXXX, XXX, XXX−XXX

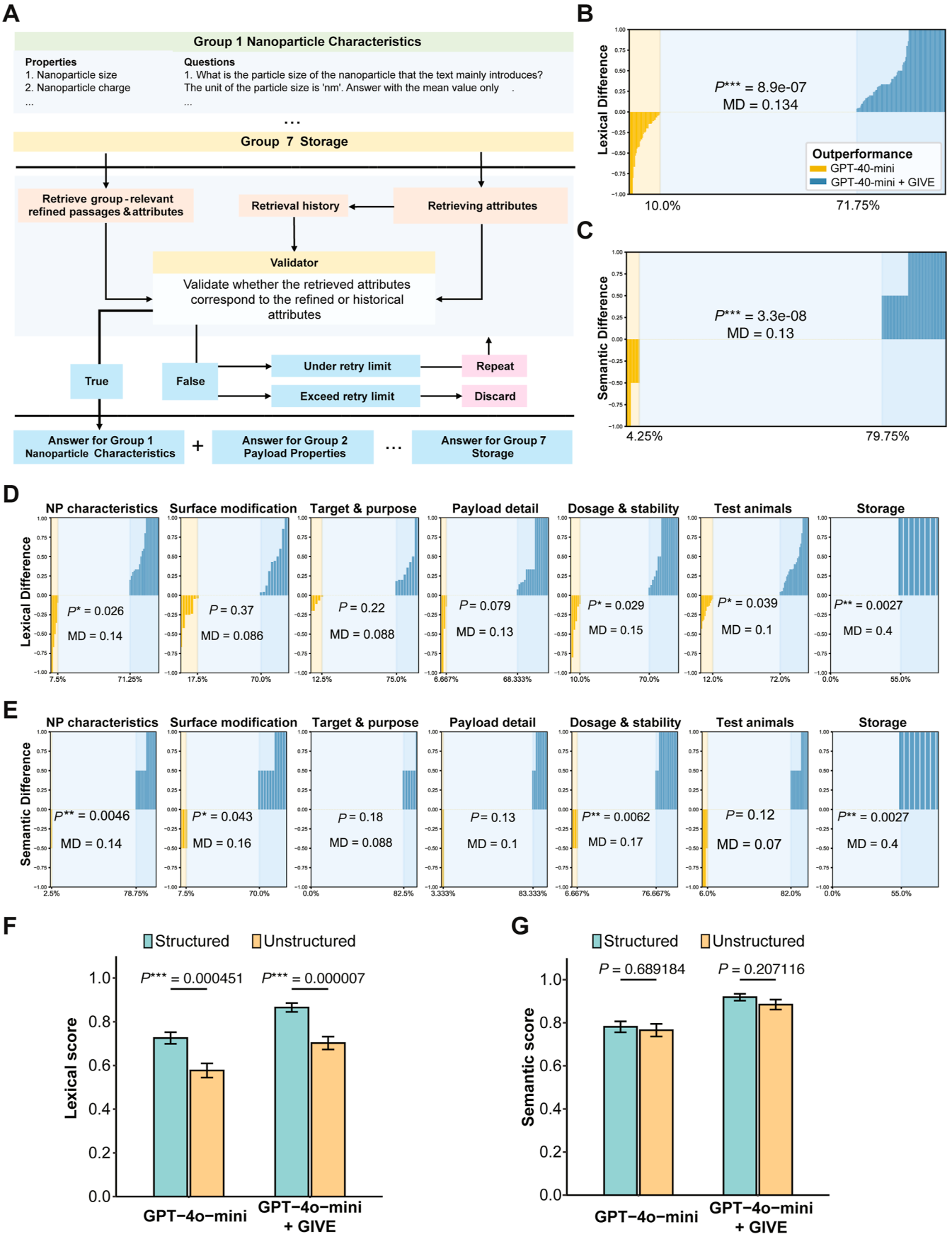


Figure 2. Extraction and evaluation of information from publications. (A) Workflow for the extraction pipeline. Attributes are grouped and then extracted and validated iteratively by comparison with a refined extraction as well as historical extraction. (B–E) Extraction comparison between the GPT-4o-mini and Grouped Iterative Validation based Information Extraction (GIVE) method. Each method was tested for

https://doi.org/10.1021/acsnano.5c03454
ACS Nano XXXX, XXX, XXX−XXX

Figure 2. continued

attribute extraction across 400 iterations. (B) Overall comparison by lexical evaluation. (C) Overall comparison by semantic evaluation. (D,E) Grouped extraction comparison by lexical evaluation (D) or semantic evaluation (E) between GPT-4o-mini and GIVE by groups. The seven groups are nanoparticle (NP) characteristics (4 attributes, n = 80), NP surface modification (2 attributes, n = 40), target and purpose (2 attributes, n = 40), payloads (3 attributes, n = 60), dosage and stability (3 attributes, n = 60), test animals (5 attributes, n = 100), and storage (1 attributes, n = 20). (F,G) Extraction comparison between structured and unstructured attributes by lexical evaluation (F) or semantic evaluation (G). Eight attributes (n = 160) were used for unstructured data in each group and twelve attributes (n = 240) for structured data in each group. Two sample paired t-tests were used in (B−E); Welch t-tests were used in (F,G). All error bars are means ± s.e.

answer is semantically similar to the previous one in the vector space. If the answers are found to be semantically similar, then the attribute is considered valid and retained. If not, the extraction process will be repeated and updated results will be validated again. However, if no consensus can be reached after several iterations, the attribute is discarded (Figure S1). This dual-verification mechanism ensures high accuracy by retaining only attributes that showed consistency with either refined answers or historical extractions.

Next, we used two approaches to assess the performance of the information extraction: an automated lexical evaluation that relied on exact character matching and a semantic evaluation that was performed manually. The lexical evaluation looks for exact matches of words, whereas the semantic evaluation examines the meaning behind the words. For example, if the original text in the article uses the term "renal" and the extraction result was "kidney", the lexical evaluation would flag it as incorrect, but sematic evaluation would accept it as the correct answer. The information extraction accuracy of the GIVE framework was 80.03% based on lexical assessment and 90.50% through semantic evaluation. This demonstrates that the GIVE method, which utilizes GPT-4o-mini as its base LLM, achieves higher accuracy and more valid extraction of information across all attributes compared to using GPT-4o-mini alone (Figures 2b−e and S2−S5).

Analysis of the extracted data and predefined attributes revealed that the data could be categorized according to the intrinsic characteristics. The extracted data were classified as two distinct types: structured and unstructured. Structured data follow a specific, deterministic format or a finite set of predefined options, enabling consistent representation and easier processing. In contrast, unstructured data lack a set format and can vary widely in structure and presentation. For example, the term "surface modification" is considered unstructured data because it includes descriptive words or phrases describing techniques applied to alter the surface properties of materials. On the other hand, the term "particle size" is structured data, as it consists of a single numeric value representing the nanoparticle's exact size. Analysis of the results showed that for unstructured data, semantic evaluation resulted in higher scores than lexical evaluation, suggesting that lexical evaluation may fail to accurately match unstructured attributes with their corresponding answers because of the undetermined format. Further analysis also supported this hypothesis: lexical evaluation tended to underperform on extracted unstructured attributes, such as "surface modification" and "Targeted site" (Figure 2f, Table S5). For those attributes, variations in phrasing of the same answer can lead to failures in lexical evaluations that rely on exact character matches. In addition, factors such as abbreviations, chemical formulas, and other nuances can also affect the lexical evaluation of unstructured attributes. Since relational databases use methods similar to lexical match for retrieval and prioritize returning higher-scoring results, unstructured attributes are often not effectively retrieved using these methods in such systems.

On the other hand, the semantic evaluation performance showed no significant difference between unstructured and structured attributes (Figure 2g). Given this observation, we designed a hybrid database by adopting different search methods for unstructured and structured attributes to eliminate this bias with regard to searching different types of attributes. We removed the lexical retrieval for unstructured attributes to mitigate the ineffectiveness of this method. As a result, a relational database was established for lexical search—retrieved structured data, and a vector database was established for semantic search—retrieved unstructured data. For each paper, unstructured attributes were concatenated, converted into embeddings, and stored in the vector database, while structured attributes were stored separately in the relational database (Table S6). This storage process enabled the efficient retrieval of both structured and unstructured attributes.

Characterizing the Extracted Information on Nanoparticles. We next analyzed the information we gathered in the database to characterize the dataset. Overall, we were able to extract the essential and critical nanoparticle attributes from >80% of the total papers (Figure 3a). We further examined and characterized the data in the database by validating the data trends with known knowledge. First, we checked the association of certain characteristics with the nanoparticle type. As expected, the parameter "lipid cholesterol ratio" appeared more often for the liposomal type of nanoparticle than for other types, especially inorganic nanoparticles (44.79% vs 2.77%, Figure 3b). The parameter "gene length" was most commonly available for nanoparticles when the payload type was nucleic acid (Figure 3b). The relationship between different attributes extracted from each article was further illustrated by chord graphing (Figure 3c). We further examined whether these trends reflect the characteristics of different nanoparticles. Compared with inorganic nanoparticles, liposomal nanoparticles had more available attributes of "lipid and cholesterol ratio" (Figure 3d). Also, the attribute "gene length" was often available when the payload was nucleic acid and was nearly absent when the payload was small molecule (Figure 3d). The overall frequency of extracted data for different types of nanoparticles and payloads, and their chances of coexisting with other parameters, is shown in Figures S6 and S7. We also examined the association between the numeric values of nanoparticle parameters and their characters. The size of nanoparticles showed different distributions based on the targeted site. For example, kidney-targeting nanoparticles were smaller than liver-targeting nanoparticles (Figure S8). The distribution of extracted attributes is summarized in Figures S9−S11. In sum, the

https://doi.org/10.1021/acsnano.5c03454
ACS Nano XXXX, XXX, XXX−XXX

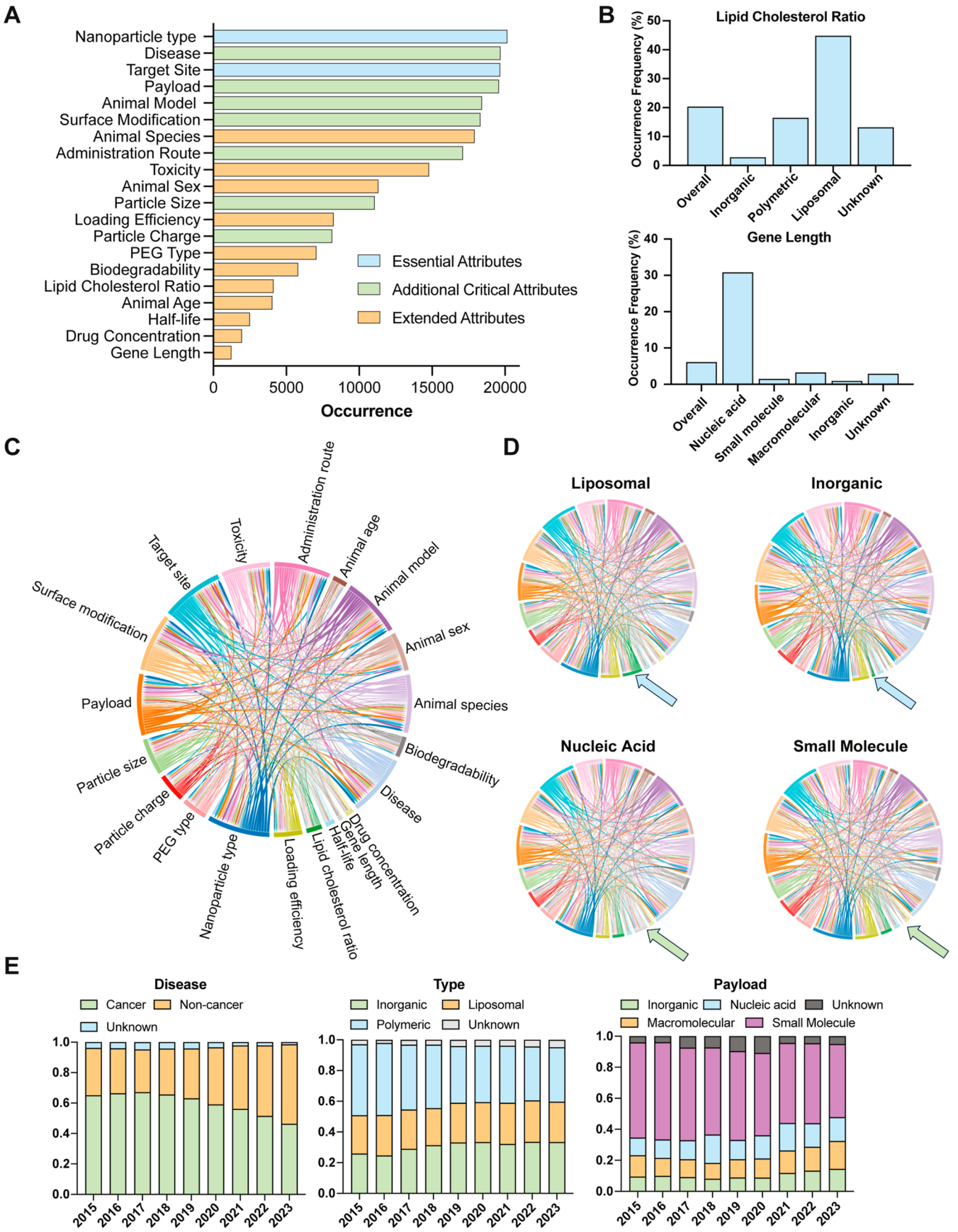


Figure 3. Characterization of the nanomedicine database. (A) Overview of the database attributes. (B) Occurrence of lipid cholesterol ratio by nanoparticle type (n = 4,127) and occurrence of gene length by payloads (n = 1,235). (C) Co-occurrence chord graph of database attributes. Each arc represents an attribute, with the width of the arc proportional to the frequency with which the attribute appears in the

https://doi.org/10.1021/acsnano.5c03454
ACS Nano XXXX, XXX, XXX−XXX

Figure 3. continued

dataset. The chords connecting the arcs signify the co-occurrence of attributes, with the chord width indicating the frequency of their joint occurrence. *n* = 20,338 independent papers. (D) Co-occurrence chord graph of database attributes for payloads. Arrows point to the main differences between graphs. (E) Distribution of key attributes in nanoparticle drug delivery research by year, illustrated for disease type (cancer, noncancer, or infection); nanoparticle type (inorganic, liposomal, or polymeric); and payload type (inorganic, nucleic acid, macromolecular, or small molecule). "Unknown" represents the proportion of articles for which the key attributes are not specified or could not be classified.

information gathered in our nanoparticle characterization database reflects known "real-world" knowledge. The extracted data also enabled us to track the publication trends. The percentage of cancer-related research in the nanomedicine field remained stable between 2015 and 2020 but decreased since 2021 in parallel with an increase in noncancer research (Figure 3e), probably because of growing interest in vaccines after the COVID-19 pandemic. The nanoparticle type in the studies remained stable among the major classifications of inorganic, liposomal, and polymeric (Figure 3e). For the payloads being delivered by the nanoparticles, we saw a slight but steady decline in small molecules and an increase in inorganic and macromolecules over time (Figure 3e).

Overview of Multiagent Question Answering. Next, to use the information in our database to achieve better responses from the LLM, we designed an application, NanoSafari, to incorporate information from the query. The answering framework enables users to use natural language inputs and receive accurate and relevant advice from reliable sources in an interactive manner. The framework uses the concept of multiagent system,[29] Chain of Thought,[30] and least-to-most prompting,[31] which function by deconstructing a complex question into a sequence of intermediate steps to be resolved independently by several agents (Figure 4a). First, the parsing agent validates whether a query is relevant to the field of nanomedicine; if the agent determines that the query is irrelevant or unreasonable, it would be rejected by returning a predefined message to prompt the user to amend the query. Otherwise, the agent deconstructs the query into fine-grained subqueries and extracts key information for further processing. Next, the searching agent retrieves information from both relational and vector databases by lexical and semantic search methods. The searching agent incorporates a structured query language (SQL) generation tool, which is responsible for generating SQL queries based on the structured output provided by the parsing agent, thereby enabling the lexical extraction of data from the relational database. A vector database search tool is also integrated, allowing the agent to perform semantic searches and retrieve conceptually similar data from the vector database. The retrieved data are ranked by the relevance to the question and citation counts and passed to the next agent, the summarizing agent, which consolidates all of the information collected from the databases (i.e., the output from the search agent) and responds to a user's query with a structured and comprehensive answer. In addition to providing a clear and organized response, the summarizing agent ensures that the relevant papers and sources are cited within the text, allowing the user to trace the information back to the original research.

NanoSafari Provides Better Results than GPT. To test the performance of NanoSafari, we generated answers to 10 nanoparticle design questions by using NanoSafari or GPT-4o-mini in parallel (Table S7). The questions were related to a variety of different nanoparticle types, payloads, and tasks. We invited postdoctoral fellows in the field of nanomedicine to blindly evaluate the randomized answers in terms of four variables: overall helpfulness, depth of knowledge, depth of details, and faithfulness. The results showed that the users preferred the answers from NanoSafari significantly more often than those from GPT-4o-mini ($P < 4.5 \times 10^{-6}$, Figure 4b). Examination of the reference(s) in the generated answers for validity revealed that on average, only about 35% of the references provided by GPT existed, whereas 100% of the references provided by NanoSafari were real (Figure 4c). Further manual evaluation confirmed the validity of the references and high relevance to the query and output. These results suggest that NanoSafari can generate more helpful and reliable answers to address the needs of researchers compared with GPT-4o-mini.

Continuous Self-Evolution of NanoSafari. Considering that NanoSafari relies on information in the database to provide relevant and accurate results, we tested if adding new papers to enrich the database could further improve the performance of NanoSafari. We designed the experiment by forcing NanoSafari to refer to a subdatabase with information extracted from 500 irrelevant papers as the control group and used another subdatabase consisting of the 500 irrelevant papers plus 500 relevant papers as the test group (Table S8). Blinded scoring showed that the addition of relevant papers significantly improved the quality of generated answers in terms of overall helpfulness, faithfulness, and richness of details (Figure 4d). This result suggested that NanoSafari had the potential for continuous evolution and prompted us to establish a platform for the research community to feed the database with the latest information. To do so, we enabled NanoSafari to allow users to upload copies of research papers for automated information extraction using GIVE. The whole extraction process is automated. With this function, members of the nanomedicine research community can contribute to the database and improve its abilities (Figure 4e).

NanoSafari Provided Better Nanoparticle Designs than GPT. Finally, we undertook experiments to validate whether the results generated by NanoSafari could help with bench work. We asked NanoSafari and GPT-4o-mini to provide design suggestions to formulate lipid nanoparticles to be loaded with green fluorescent protein (GFP) mRNA. Both outputs suggested using cationic lipids and PEGylation but showed differences in lipid type and ratios of reagents. Notably, GPT suggested using 10% DSPE-PEG2000, which is out of the usual range according to our experience and published work. GPT also suggested using A18-Iso5p-P18 as the cationic lipid, for which we could not find ordering information and used A18-Iso5-2DC18 instead. In contrast, NanoSafari suggested DLin-MC3-DMA, which was readily available. NanoSafari further suggested using a 1.5% ratio of DSPE-PEG2000, which was within a reasonable range and was consistent with the percentage of PEG in the provided refs 32 and 33. We then formulated the lipid nanoparticles by using

https://doi.org/10.1021/acsnano.5c03454
ACS Nano XXXX, XXX, XXX−XXX

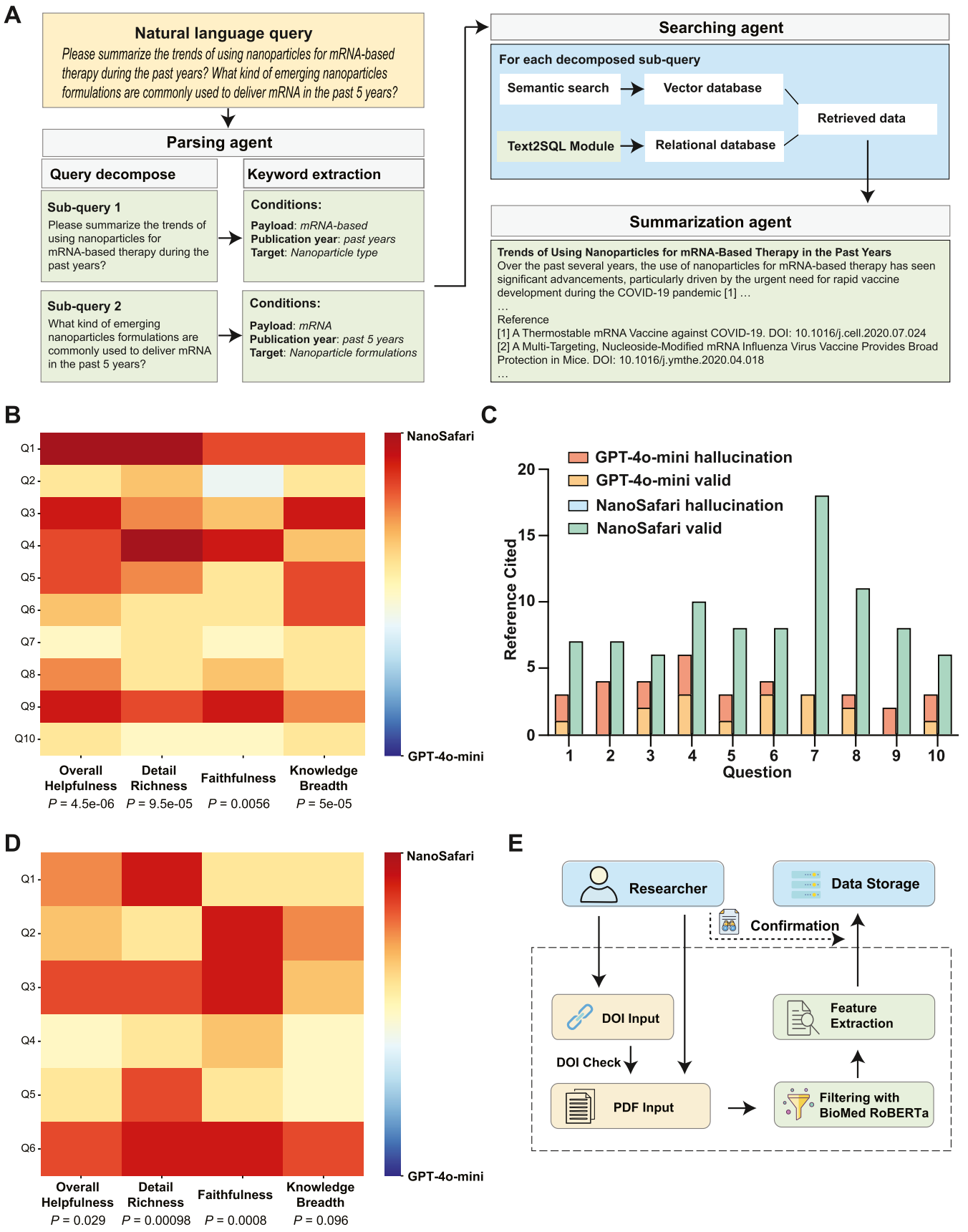


Figure 4. Database support enhances content generation. (A) Workflow for question-answering framework. The framework consists of three main components: a parsing agent that decomposes queries into fine-grained questions; a searching agent that retrieves relevant information from databases; and a summarization agent that compiles the retrieved data into a comprehensive answer. (B) Comparison of rater preferences between GPT and NanoSafari in an A/B Test across 10 questions. Participants (n = 6) evaluated responses on 4 dimensions

https://doi.org/10.1021/acsnano.5c03454
ACS Nano XXXX, XXX, XXX−XXX

Figure 4. continued

(general helpfulness, detail richness, faithfulness, and knowledge breadth) for each question. (C) Comparison of valid reference ratio between GPT-4o-mini and NanoSafari. *n* = 10 answers for each group. (D) Comparison of rater preferences between the uncorrelated data-supporting system and data-supporting system supplemented with highly correlated data in an A/B Test across 6 questions. Participants (*n* = 8) evaluated responses on 4 dimensions for each question. (E) Workflow for the crowdsourcing pipeline. Users submit either DOI or PDF inputs, which are validated through DOI checks. PDF inputs are filtered by BioMed RoBERTa, followed by manual validation. Validated documents undergo feature extraction, and the extracted data are stored in the file storage system. Single sample *t* tests were used in (B,D). Heatmaps were used in (B,D) to illustrate the preferences of participants when presented with two answer options. The color intensity of each cell corresponds to the number of participants who favored the answer associated with that cell. Detailed keys are provided to interpret the color scale in relation to participant counts.

the reagents and ratios provided by NanoSafari and GPT and loaded them with GFP mRNA. The size distribution of the nanoparticles generated by the NanoSafari formula was more homogeneous than the GPT formula (Figure 5a). The NanoSafari nanoparticles also had better mRNA-loading efficiency (Figure 5b) and achieved more GFP expression after a 24 h incubation with mouse embryonic fibroblasts (MEFs) than did the GPT nanoparticles (Figure 5c). We also tested those lipid nanoparticles for cancer vaccine applications. When loaded with OVA mRNA, the lipid nanoparticles made by the NanoSafari formula resulted in more SIINFEKL peptide presented by bone-marrow-derived dendritic cells compared with the GPT counterpart (Figure 5d,e). Next, we asked NanoSafari and GPT about size selection for polystyrene nanoparticles intended for lung delivery. NanoSafari suggested a 10−50 nm size for effective diffusion through the alveolar-capillary membrane and provided relevant citations discussing nanoparticle lung delivery.[34,35] On the contrary, GPT suggested using a 100−200 nm size and provided a citation. Upon examination, GPT hallucinated the title and conclusion of the cited paper by claiming 100−150 nm is the ideal size for lung delivery, and even provided a real Digital Object Identifier (DOI) of a paper studying spintronics spin-based quantum information processing,[36] which was completely irrelevant. To further test the suggestions, we injected same doses of 25 and 100 nm fluorescent PEGylated polystyrene nanoparticles to mice via the tail vein. Ex vivo imaging showed that the NanoSafari-suggested 25 nm nanoparticles had more lung delivery than the GPT-suggested 100 nm nanoparticle (Figure 5f,g). Together, these results demonstrate the ability of NanoSafari to provide reliable suggestions for a variety of different nanoengineering applications.

DISCUSSION

LLMs are emerging platforms to assist in nanoengineering. Among several recent applications of LLMs introduced in biomedical research are their use in training models to predict the transfection efficiency of lipid nanoparticles.[37] Another LLM-based application has been developed to assist with experimental designs for gene editing.[38] To train the AI models, these studies use relatively "clean" experimental data, which are uniform and consistent in terms of the format from experimental results. This approach has the advantage of data uniformity and consistency but could be limited by the scope of the original experiments from which the data were generated. In addition, updates to applications trained by experimental data rely on the input of new data, which could be delayed relative to more rapid developments in some scientific fields. The key innovation of our study is that it relies on the analysis of the natural language of the published literature to automatically extract contextual information from both published research and community-fed new research, which makes it capable of self-evolving for continuous enhancement.

Humans have used language as a carrier for experience and knowledge for thousands of years. In the foreseeable future, natural language will still be the medium for exchanging scientific discoveries and ideas. A group of experts in nanomedicine proposed standards for reporting nanostudies nearly a decade ago.[39] However, variations among journal formats and preferences of authors still complicate the process for summarizing large amounts of data in a centralized database. Our study showcases a feasible approach to garner vast amounts of data in a research area by using natural language processing to break down some of the barriers between human natural language and computational algorisms, and we demonstrated that this effort could significantly enhance the capability and reliability of AI to assist scientific research. While human creativity and rigor will remain the primary drivers of scientific discovery for the foreseeable future, empowering AI to efficiently generate reliable insights from the published literature will undoubtedly accelerate this process. Considering the large amount of published literature and new research findings being published each year, our approach, with increasingly sophisticated abilities, holds great potential to be applied to broader areas of research. Future work may include additional information to be extracted from the literature, such as experimental steps, to further enrich the nanoparticle database and therefore enhance the capability of AI to assist in designing experiments. However, the scope and depth of extracted data rely on the information reported by the paper. The ongoing efforts to enhance the transparency and reporting standards of scientific papers can potentiate more enriched and reliable information extraction from future publications.

CONCLUSION

Our application provides a novel and robust platform to assist in designing nanoparticles for biomedical engineering, especially for the purpose of drug delivery. A key advance of this application is that it extracts information from published work to generate a database with accurate and reliable information as an index to substantially enhance the reliability of LLMs. The ability of the NanoSafari platform to continuously improve by integrating new data from publications also ensures that it can keep pace with the latest discoveries in the field. These features make NanoSafari a reliable "copilot"′′ to assist with and accelerate research in biomedical nanoengineering.

https://doi.org/10.1021/acsnano.5c03454
ACS Nano XXXX, XXX, XXX−XXX

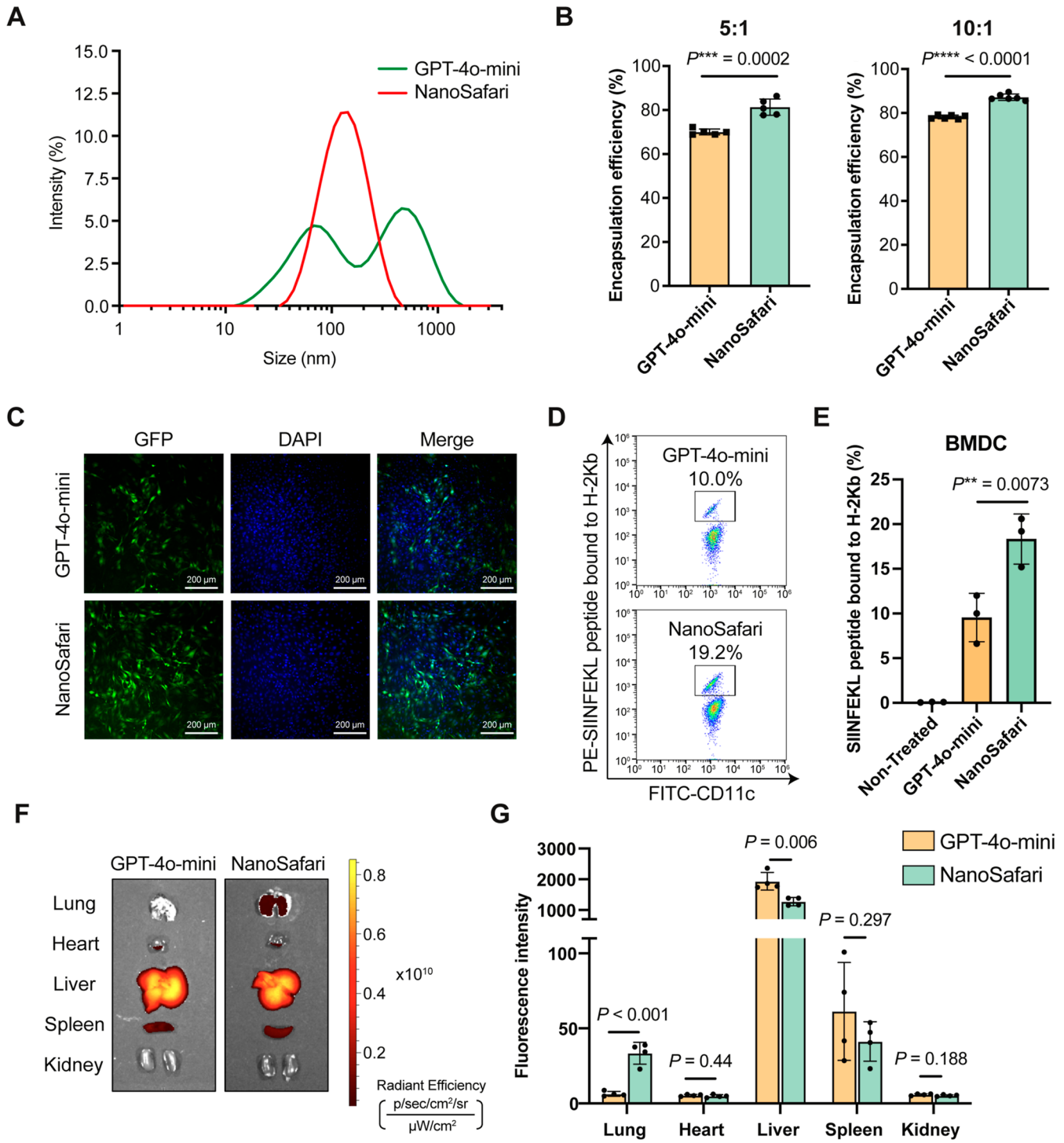


Figure 5. NanoSafari effectively supported bench experiments. (A) Size distribution of LNPs generated using formulations provided by GPT-4o-mini or NanoSafari. (B) mRNA loading efficiency of two different LNPs. The mass ratio of total lipid components (including cholesterol) to mRNA was set at either 5:1 or 10:1. n = 5 biological independent replicates. Unpaired two-sided t-test. Error bars are means ± s.d. (C) Representative cell transfection results using two different LNP formulations. (D) Representative flow cytometry result of BMDC transfected with OVA mRNA-LNP. (E) Summarized result of OVA peptide presented by BMDC. n = 3. One-way ANOVA with correction for multiple comparisons. Error bars are means ± s.d. (F) Representative image of organs after 4 h of nanoparticle injection. (G) Summarized results of nanoparticle biodistribution. n = 4 mice each group. Multiple unpaired t-tests. Error bars are means ± s.d.

METHODS

Human Evaluation of AI Outputs. Human participants evaluated two AI-generated responses to a single question. The responses from the different AI sources were randomized and the participants were unaware of the origin of the answers to ensure blinded test conditions. All participants held a PhD in biomedical or biological science or an MD, and all had at least 2 years of research experience in nanomedicine or drug delivery. The participants were tasked with judging the superiority of one answer over the other across four key dimensions: (1) overall helpfulness, (2) breadth of knowledge, (3) details and depth, and (4) faithfulness ("Faithfulness" is described in further detail in the paragraph on "Lexical Evaluation

https://doi.org/10.1021/acsnano.5c03454
ACS Nano XXXX, XXX, XXX−XXX

Metrics"). The participants were instructed to choose from one of three responses for each category: (1) source A > source B, (2) A = B, or (3) A < B. Correspondingly, the answers would be translated into numerical values of 1, 0, and −1, respectively, for statistical analysis.

Literature Search. Publications of interest were obtained from the Web of Science (clarivate.com) by using the following query: ((TS = nanoparticle) OR (TS = nano particle) OR (TS = nanomaterial) OR (TS = nano material) OR (TS = nanoscale material) OR (TS = nanocarrier) OR (TS = nano carrier) OR (TS = nanomedicine) OR (TS = nano medicine)) AND ((TS = drug delivery) OR (TS = chemotherapy) OR (TS = chemotherapy delivery) OR (TS = inhibitor delivery) OR (TS = agonist delivery) OR (TS = gene delivery) OR (TS = peptide delivery) OR (TS = antibody delivery) OR (TS = mRNA delivery) OR (TS = gene delivery) OR (TS = siRNA delivery) OR (TS = mRNA therapy) OR (TS = therapy delivery)) AND PY = 2015−2023. PDFs of the selected papers were downloaded by the study team from the journal Web sites, through either institutional subscriptions or open access. Some papers were not included in the final analysis owing to subscription limits.

Attributes to Extract and Datatypes. Twenty attributes were extracted from each nanomedicine paper and ranked as essential, important, or extensive. Essential attributes were mandatory for inclusion; important attributes were valuable but not critical; and extensive attributes were supplementary. Attributes were categorized according to their datatype, classified as TEXT, FLOAT, ENUM, or BOOLEAN. Here, TEXT data contain free-form text; FLOAT data represent numerical data with decimals; ENUM attributes represent a limited set of predefined data; and BOOLEAN data are True or False only.

Filtering Papers with the RoBERTa Classifier. To filter out off-topic papers, we used the Robustly Optimized BERT Pretraining model (RoBERTa),[40] which is based on the architecture of BERT.[41] RoBERTa uses an encoder-only structure from the Transformer model, offering a more "lightweight" alternative to LLMs while demonstrating strong performance across various natural language processing tasks. In this case, our focus was on papers that explicitly discuss the process of nanoparticle drug delivery.

A total of 500 papers were manually annotated as either positive or negative, with 246 papers validated as positive and the remaining papers classified as negative. Once the filtering process was complete, the abstracts of both positive and negative papers were then used as training data for RoBERTa to perform the classification tasks.

Lexical Evaluation Metrics. The metrics of the lexical evaluation were considered in two categories: correctness and faithfulness. A response was deemed correct if it accurately met the user's information needs. For instance, for the question what is the animal species of the experiment, the model must correctly identify mouse as the species. Although the response may include additional details, such as the animal's sex or age, the evaluation focuses solely on the portion directly relevant to the user's query. "Correctness" in quality assurance is assessed by comparing model responses against concise human-annotated gold standards with recall, which is calculated as the proportion of tokens in the reference answer that are present in the model response. The model's responses often include additional information beyond the user's specific request. For example, for the question what is the animal species of the experiment, the model might respond with details such as the age—6−11 weeks mouse. Evaluating this supporting information can be challenging without human annotation. Consequently, we focused on a more limited objective: faithfulness. A faithful response should be fully grounded in the provided knowledge. Faithfulness in quality assurance is assessed by comparing model responses against relevant knowledge, which is longer text snippets from the source, with K-precision (knowledge-precision), calculated as the proportion of tokens in the model response that appear in the knowledge snippet. We comprehensively considered the two categories by taking the harmonic mean of both as the criterion for lexical evaluation, defined by the following formula

$$H = \frac{2}{\text{recall}^{-1} + \text{Kprecision}^{-1}}$$

Semantic Evaluation Metrics. Semantic performance was assessed through manual scoring by comparing the extracted information to human-annotated gold standard answers to validate consistency. Raters were tasked with determining whether a given answer was (1) fully correct, (2) partially correct, or (3) incorrect, assigning scores of 1, 0.5, or 0, respectively.

Embedding Model and Semantic Similarity. "Embeddings" were generated by using the text-embedding-3-large model from OpenAI (2024). This embedding model processes text inputs and maps them into high-dimensional vector spaces, effectively capturing their semantic meanings. To determine the semantic similarity between two pieces of text, cosine similarity is computed based on their corresponding embeddings. Semantic search is conducted by retrieving the embedding of textual data that exhibit the highest cosine similarity to the embedding of a given input text.

Text2SQL Module. The Text2SQL agent, which uses GPT-4o's capability to achieve function calling, enables the model to autonomously decide whether to request the database. If the model determines that the query is seeking information from the database, it will then formulate the necessary SQL query based on the keywords it has extracted. After generating the SQL query, the module proceeds with postprocessing by using regular expression.[42]

Establishment of the Subdatabase. The user's query facilitates a search within the vector database, sorting all articles by their relevance to the query as indicated by the search results. The control group subdatabase consists of the 500 least relevant articles, identified by their lowest ranking in the search order. In contrast, the test group subdatabase is formed by supplementing the control group with the 500 most relevant articles, which are the highest ranked in the search results.

Crowdsourcing. The NanoSafari system can be accessed at https://nanosafari.com/. The platform is open to the public by adopting a crowdsourcing approach. Users are invited to upload scholarly literature to our database. The system systematically extracts pertinent attributes and integrates the paper into the database. This approach leverages the collective knowledge and expertise of the community, ensuring that the database remains comprehensive and up to date. The expert review process guarantees the quality and relevance of the literature, whereas automated attribute extraction facilitates efficient organization and retrieval of information. By involving a diverse range of contributors, the system aims to democratize access to scientific knowledge and foster collaboration among researchers. The crowdsourcing model not only enhances the breadth of the database but also encourages active participation and engagement from the academic community. As more users contribute to and review the contributed literature, the system continuously evolves, adapting to emerging trends and new areas of research. This dynamic and interactive platform thus serves as a valuable resource for researchers, providing a rich repository of validated and well-organized scholarly works.

Preparations of mRNA-LNPs. The LNP formulations in this study were prepared using two different formulations, referred to as "GPT-4o-mini" and "NanoSafari", each employing distinct lipid compositions. Formulation of LNP followed the suggested formula by each model. Each model generated outputs to the same prompt 20 times independently, and the results were consistent. For the NanoSafari, the ionizable lipid DLin-MC3-DMA, helper lipid 1,2-distearoyl-*sn*-glycero-3-phosphocholine (1,2-DSPC), cholesterol, and PEG lipid DMG-PEG2000 were dissolved in ethanol at a molar ratio of 50:10:38.5:1.5. For GPT-4o-mini, the ionizable lipids A18-Iso5-2DC18, 1,2-DSPC, cholesterol, and DMG-PEG2000 were dissolved in ethanol at a molar ratio of 50:30:10:10. EGFP-mRNA was prepared in a 25 mM sodium acetate buffer at pH 4.0. Nanoparticles were formed by rapidly injecting the lipid solution into the mRNA-containing aqueous phase while continuously stirring. The mass ratio of total lipid components (including cholesterol) to mRNA was set at either 5:1 or 10:1. The resulting LNP suspension was dialyzed against

https://doi.org/10.1021/acsnano.5c03454
ACS Nano XXXX, XXX, XXX−XXX

a 100-fold volume of 25 mM sodium acetate buffer (pH 4.0) to eliminate residual ethanol.

Characterization of mRNA-LNP Formulations. Following dialysis, the size of the LNP formulations was measured using dynamic light scattering (DLS) with a Malvern Zetasizer ZS90. The encapsulation efficiency of the LNPs was assessed using the Quant-iT RiboGreen Assay (Thermo Fisher Scientific, Cat# R11490). Briefly, LNPs were treated with 0.5% (w/v) Triton X-100 to disrupt the LNP structure and release encapsulated mRNA. Both treated and untreated LNPs were diluted to a concentration below 1 μg of mRNA mL^{-1} and incubated with an equal volume of RiboGreen assay solution (200-fold dilution). Standard curves were generated using free mRNA solutions with or without 0.5% (w/v) Triton X-100, spanning a concentration range of 0.1 to 1.0 μg of mRNA mL^{-1}. The concentrations of free and total mRNA in the formulations were quantified by measuring fluorescence (excitation: 480 nm; emission: 520 nm) and comparing the values to the corresponding standard curves. To test the transfection efficiency, the mRNA-LNP was incubated with MEF cells or bone-marrow-derived dendritic cells. Fluorescence imaging or flow cytometry was used to test protein expression after 24 h.

Animal Experiments. Wild-type C57BL6 mice were purchased from Jackson Laboratory. Male mice 8−12 weeks old were used. The mice were housed in a specific-pathogen-free environment in a 12 h light/12 h dark cycle with 50% humidity and 22 °C temperature. Fluorescent PEGylated polystyrene nanoparticles were prepared as previously described[43] and were injected into mice via the tail vein with a dose of 10^{12} particles/mouse. After 4 h of injection, the mice were euthanized, and the organs were imaged to quantify fluorescence intensity by the IVIS imaging system.

Data Analysis. Two-sided paired *t*-tests were used to compare differences between two matchable groups, and Welch *t*-tests were used to compare differences between two mismatched data groups. The statistical methods, meaning of error bars, and numbers of independent replicates of each figure are all given in the corresponding figure legends. Significance levels are labeled with asterisks: $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. Statistical analyses were done with IBM SPSS Statistics 25.0, Python 3.9, or R 4.2.3 software.

ASSOCIATED CONTENT

Data Availability Statement

All data used to generate the results in this paper are available in the manuscript, and Supporting Information. Additional data or information can be requested by contacting the corresponding authors (Z.W., B.Y.S.K, and W.J.). The codes used in the study have been deposited into GitHub and are accessible at https://github.com/ytten/NanoSafari.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsnano.5c03454.

Workflow of information extraction from publications by GIVE; performance comparison between GPT and GPT + GIVE by lexical and human evaluations for each group of attributes; overall performance comparison between GPT and GPT + GIVE by lexical and human evaluations; comparison of textual and nontextual attributes extracted by GPT vs GPT + GIVE; extraction comparison between GPT-4o-mini and GPT-4o-mini + GIVE; co-occurrence chord graph of database attributes; heatmap of database attributes; particle size distributions across multiple target sites; particle charge distributions across multiple target sites; distributions of various experimental parameters across multiple categories; distributions of various experimental parameters across multiple categories; keywords used to search papers about nanoparticle drug delivery; attributes grouped by importance; attributes grouped by category; questions about nanoparticle drug delivery; regression of extraction performance; schema for the nanoparticle drug delivery database; results of GPT-4o-mini and NanoSafari responses to professional questions; and results of NanoSafari with different database responses to professional questions (PDF)

AUTHOR INFORMATION

Corresponding Authors

Zikai Wang − *SigmaSafari Project, Shanghai Artificial Intelligence Research Institute Co., Ltd., Shanghai 201109, China; Xiangfu Laboratory, Jiashan 314100, China*; Email: wangzikai@sairi.com.cn

Betty Y. S. Kim − *Department of Neurosurgery, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, United States*; Email: bykim@mdanderson.org

Wen Jiang − *Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, United States*; orcid.org/0000-0001-9154-633X; Email: wjiang4@mdanderson.org

Authors

Yifan Wang − *Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, United States*; orcid.org/0000-0001-7387-0654

Haitao Song − *SigmaSafari Project, Shanghai Artificial Intelligence Research Institute Co., Ltd., Shanghai 201109, China; Xiangfu Laboratory, Jiashan 314100, China*; orcid.org/0000-0003-2113-2224

Yue Teng − *SigmaSafari Project, Shanghai Artificial Intelligence Research Institute Co., Ltd., Shanghai 201109, China; Xiangfu Laboratory, Jiashan 314100, China*

Guan Huang − *SigmaSafari Project, Shanghai Artificial Intelligence Research Institute Co., Ltd., Shanghai 201109, China; Xiangfu Laboratory, Jiashan 314100, China*

Jingzhe Qian − *SigmaSafari Project, Shanghai Artificial Intelligence Research Institute Co., Ltd., Shanghai 201109, China; Xiangfu Laboratory, Jiashan 314100, China*

Hongyu Wang − *SigmaSafari Project, Shanghai Artificial Intelligence Research Institute Co., Ltd., Shanghai 201109, China; Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong SAR 999077, China*

Shiyan Dong − *Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, United States*

JongHoon Ha − *Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, United States*

Yifan Ma − *Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, United States*; orcid.org/0000-0002-9371-4365

Mengyu Chang − *Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, United States*

Seong Dong Jeong − *Department of Neurosurgery, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, United States*

https://doi.org/10.1021/acsnano.5c03454
ACS Nano XXXX, XXX, XXX−XXX

Weiye Deng − *Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, United States*
Benjamin R. Schrank − *Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, United States*
Adam Grippin − *Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, United States*
Annette Wu − *Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, United States*
Jared L. Edwards − *Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, United States;* orcid.org/0009-0008-7604-3441
Yixiang Zhang − *SigmaSafari Project, Shanghai Artificial Intelligence Research Institute Co., Ltd., Shanghai 201109, China*
Yuanyuan Lin − *SigmaSafari Project, Shanghai Artificial Intelligence Research Institute Co., Ltd., Shanghai 201109, China; Xiangfu Laboratory, Jiashan 314100, China*
Wilson Poon − *Department of Metallurgical, Materials and Biomedical Engineering, The University of Texas El Paso, El Paso, Texas 79968, United States; Border Biomedical Research Center, The University of Texas at El Paso, El Paso, Texas 79968, United States;* orcid.org/0000-0003-2192-3077
Stefan Wilhelm − *Stephenson School of Biomedical Engineering, University of Oklahoma, Norman, Oklahoma 73019, United States; Institute for Biomedical Engineering, Science, and Technology (IBEST), Norman, Oklahoma 73019, United States; Stephenson Cancer Center, Oklahoma City, Oklahoma 73104, United States; Harold Hamm Diabetes Center, University of Oklahoma, Oklahoma City, Oklahoma 73104, United States;* orcid.org/0000-0003-2167-6221
Ye Bi − *The Public Experimental Center, Changchun University of Chinese Medicine, Changchun 130117, China*
Lesheng Teng − *School of Life Science, Jilin University, Changchun 130012, China;* orcid.org/0000-0003-1623-5384

Complete contact information is available at: https://pubs.acs.org/10.1021/acsnano.5c03454

Author Contributions

‖‖Y.W., H.S., Y.T., G.H., J.Q., and H.W. contributed equally. W.J., B.Y.S.K., Z.W., and Y.W. conceived the project and were responsible for all phases of the research; Y.W., H.S., Y.T., G.H., J.Q., H.W., S.D., Y.L., Y.B., L.T., and Z.W. developed the methodology; H.S., Z.W., G.H., Y.T., and J.Q. developed the code and database; Y.W., H.S., Y.T., G.H., J.Q., H.W., S.D., Y.B., J.H., M.C., S.D. J., W.D., Y.M., J.L.E., Y.Z., and A.G. performed the experiments and collected the data; Y.W., H.S., Y.T., G.H., J.Q., H.W., S.D., W.D., B.R.S., A.W., L.T., W.P., S.W., B.Y.S.K., and W.J. analyzed and interpreted the data; Y.W., Z.W., B.Y.S.K., and W.J. drafted the manuscript; all authors edited and approved the final manuscript.

Notes

The human-participant aspects of this study were reviewed and exempted by the Institutional Review Board of MD Anderson Cancer Center (submission number 2024-1133). All participants were introduced to the background and purpose of the study and signed a written informed consent to participate. The animal experiments were reviewed and approved by the Institution Animal Care and Use Committee of MD Anderson Cancer Center (protocol number 00002163).
The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank C.F. Wogan of the Division of Radiation Oncology at MD Anderson Cancer Center for editing an early version of this manuscript. The authors thank K. Zhu of the Department of Biostatistics, University of Washington, for suggestions in study design.

REFERENCES

(1) Kim, B. Y.; Rutka, J. T.; Chan, W. C. Nanomedicine. *N. Engl. J. Med.* **2010**, *363* (25), 2434−2443.
(2) Lee, D.; Huntoon, K.; Lux, J.; Kim, B. Y. S.; Jiang, W. Engineering nanomaterial physical characteristics for cancer immunotherapy. *Nat. Rev. Bioeng.* **2023**, *1* (7), 499−517.
(3) Shi, J. J.; Kantoff, P. W.; Wooster, R.; Farokhzad, O. C. Cancer nanomedicine: progress, challenges and opportunities. *Nat. Rev. Cancer* **2017**, *17* (1), 20−37.
(4) Jiang, W.; Kim, B. Y.; Rutka, J. T.; Chan, W. C. Nanoparticle-mediated cellular response is size-dependent. *Nat. Nanotechnol.* **2008**, *3* (3), 145−150.
(5) Jiang, W.; von Roemeling, C. A.; Chen, Y. X.; Qie, Y. Q.; Liu, X. J.; Chen, J. Z.; Kim, B. Y. S. Designing nanomedicine for immuno-oncology. *Nat. Biomed. Eng.* **2017**, *1*, 0029.
(6) Suk, J. S.; Xu, Q.; Kim, N.; Hanes, J.; Ensign, L. M. PEGylation as a strategy for improving nanoparticle-based drug and gene delivery. *Adv. Drug Delivery Rev.* **2016**, *99* (Pt A), 28−51.
(7) Barenholz, Y. Doxil(R)–the first FDA-approved nano-drug: lessons learned. *J. Controlled Release* **2012**, *160* (2), 117−134.
(8) Yardley, D. A. nab-Paclitaxel mechanisms of action and delivery. *J. Controlled Release* **2013**, *170* (3), 365−372.
(9) Friedrichs, S.; Bowman, D. M. COVID-19 may become nanomedicine's finest hour yet COMMENT. *Nat. Nanotechnol.* **2021**, *16* (4), 362−364.
(10) Blanco, E.; Shen, H.; Ferrari, M. Principles of nanoparticle design for overcoming biological barriers to drug delivery. *Nat. Biotechnol.* **2015**, *33* (9), 941−951.
(11) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583−589.
(12) Shen, W. X.; Liu, Y.; Chen, Y.; Zeng, X.; Tan, Y.; Jiang, Y. Y.; Chen, Y. Z. AggMapNet: enhanced and explainable low-sample omics deep learning with feature-aggregated multi-channel networks. *Nucleic Acids Res.* **2022**, *50* (8), No. e45.
(13) Zhu, M.; Zhuang, J.; Li, Z.; Liu, Q.; Zhao, R.; Gao, Z.; Midgley, A. C.; Qi, T.; Tian, J.; Zhang, Z.; et al. Machine-learning-assisted single-vessel analysis of nanoparticle permeability in tumour vasculatures. *Nat. Nanotechnol.* **2023**, *18* (6), 657−666.
(14) Lutz, I. D.; Wang, S.; Norn, C.; Courbet, A.; Borst, A. J.; Zhao, Y. T.; Dosey, A.; Cao, L.; Xu, J.; Leaf, E. M.; et al. Top-down design of protein architectures with reinforcement learning. *Science* **2023**, *380* (6642), 266−273.
(15) May, J. N.; Moss, J. I.; Mueller, F.; Golombek, S. K.; Biancacci, I.; Rizzo, L.; Elshafei, A. S.; Gremse, F.; Pola, R.; Pechar, M.; et al. Histopathological biomarkers for predicting the tumour accumulation of nanomedicines. *Nat. Biomed. Eng.* **2024**, *8*, 1366.
(16) Wang, Y.; Schrank, B. R.; Jiang, W.; Kim, B. Y. S. Learning what keeps nanomedicines in tumours. *Nat. Biomed. Eng.* **2024**, *8*, 1330.
(17) Wang, Y.; Xiao, Z.; Wang, Z.; Lee, D.; Ma, Y.; Wilhelm, S.; Wang, H.; Kim, B. Y. S.; Jiang, W. Multi-omics approaches to

decipher the interactions of nanoparticles and biological systems. *Nat. Rev. Bioeng.* **2025**, *3*, 333.
(18) Xu, Y.; Ma, S.; Cui, H.; Chen, J.; Xu, S.; Gong, F.; Golubovic, A.; Zhou, M.; Wang, K. C.; Varley, A.; et al. AGILE platform: a deep learning powered approach to accelerate LNP development for mRNA delivery. *Nat. Commun.* **2024**, *15* (1), 6305.
(19) Anderson, D.; Witten, J. Using artificial intelligence to develop gene therapy for the lungs. *Nat. Biotechnol.* **2024**. DOI: 10.1038/s41587-024-02491-x
(20) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, 1706.03762.
(21) Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. Large language models encode clinical knowledge. *Nature* **2023**, *620* (7972), 172−180.
(22) Kohane, I. S. Compared with What? Measuring AI against the Health Care We Have. *N. Engl. J. Med.* **2024**, *391*, 1564.
(23) Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A. Training language models to follow instructions with human feedback. In *Advances in neural information processing systems*, 2022; Vol. *35*, pp 27730−27744.
(24) McCoy, L. G.; Manrai, A. K.; Rodman, A. Large Language Models and the Degradation of the Medical Record. *N. Engl. J. Med.* **2024**, *391*, 1561.
(25) Peng, C.; Yang, X.; Chen, A.; Smith, K. E.; PourNejatian, N.; Costa, A. B.; Martin, C.; Flores, M. G.; Zhang, Y.; Magoc, T.; et al. A study of generative large language model for medical research and healthcare. *npj Digit. Med.* **2023**, *6* (1), 210.
(26) Li, Y.; Li, Z.; Zhang, K.; Dan, R.; Jiang, S.; Zhang, Y. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. *Cureus* **2023**, *15* (6), No. e40895.
(27) Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; Liu, T. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *arXiv* **2024**, 2311.05232.
(28) Shinn, N.; Cassano, F.; Berman, E.; Gopinath, A.; Narasimhan, K.; Yao, S. Reflexion: Language Agents with Verbal Reinforcement Learning. *arXiv* **2023**, 2303.11366.
(29) Talebirad, Y.; Nadiri, A. Multi-Agent Collaboration: Harnessing the Power of Intelligent LLM Agents. *arXiv* **2023**, 2306.03314.
(30) Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv* **2023**, 2201.11903.
(31) Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Cui, C.; Bousquet, O.; Le, Q.; Chi, E. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. *arXiv* **2023**, 2205.10625.
(32) Munson, M. J.; O'Driscoll, G.; Silva, A. M.; Lazaro-Ibanez, E.; Gallud, A.; Wilson, J. T.; Collen, A.; Esbjorner, E. K.; Sabirsh, A. A high-throughput Galectin-9 imaging assay for quantifying nanoparticle uptake, endosomal escape and functional RNA delivery. *Commun. Biol.* **2021**, *4* (1), 211.
(33) Carrasco, M. J.; Alishetty, S.; Alameh, M. G.; Said, H.; Wright, L.; Paige, M.; Soliman, O.; Weissman, D.; Cleveland, T. E. t.; Grishaev, A.; Buschmann, M. D. Ionization and structural properties of mRNA lipid nanoparticles influence expression in intramuscular and intravascular administration. *Commun. Biol.* **2021**, *4* (1), 956.
(34) Zhong, Q.; Merkel, O. M.; Reineke, J. J.; da Rocha, S. R. Effect of the Route of Administration and PEGylation of Poly(amidoamine) Dendrimers on Their Systemic and Lung Cellular Biodistribution. *Mol. Pharmaceutics* **2016**, *13* (6), 1866−1878.
(35) Li, C.; Guan, H.; Li, Z.; Wang, F.; Wu, J.; Zhang, B. Study on different particle sizes of DOX-loaded mixed micelles for cancer therapy. *Colloids Surf., B* **2020**, *196*, 111303.
(36) Kanai, Y.; Deacon, R. S.; Takahashi, S.; Oiwa, A.; Yoshida, K.; Shibata, K.; Hirakawa, K.; Tokura, Y.; Tarucha, S. Electrically tuned spin-orbit interaction in an InAs self-assembled quantum dot. *Nat. Nanotechnol.* **2011**, *6* (8), 511−516.
(37) Moayedpour, S.; Broadbent, J.; Riahi, S.; Bailey, M.; V Thu, H.; Dobchev, D.; Balsubramani, A.; N D Santos, R.; Kogler-Anele, L.; Corrochano-Navarro, A.; et al. Representations of lipid nanoparticles using large language models for transfection efficiency prediction. *Bioinformatics* **2024**, *40* (7), btae342.
(38) Huang, K.; Qu, Y.; Cousins, H.; Johnson, W. A.; Yin, D.; Shah, M.; Zhou, D.; Altman, R.; Wang, M.; Cong, L. CRISPR-GPT: An LLM Agent for Automated Design of Gene-Editing Experiments. *arXiv* **2024**, 2404.18021.
(39) Faria, M.; Bjornmalm, M.; Thurecht, K. J.; Kent, S. J.; Parton, R. G.; Kavallaris, M.; Johnston, A. P. R.; Gooding, J. J.; Corrie, S. R.; Boyd, B. J.; et al. Minimum information reporting in bio-nano experimental literature. *Nat. Nanotechnol.* **2018**, *13* (9), 777−785.
(40) Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, 1907.11692.
(41) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, 1810.04805.
(42) Thompson, K. Programming Techniques: Regular expression search algorithm. *Commun. ACM* **1968**, *11* (6), 419−422.
(43) Qie, Y.; Yuan, H.; von Roemeling, C. A.; Chen, Y.; Liu, X.; Shih, K. D.; Knight, J. A.; Tun, H. W.; Wharen, R. E.; Jiang, W.; Kim, B. Y. Surface modification of nanoparticles enables selective evasion of phagocytic clearance by distinct macrophage phenotypes. *Sci. Rep.* **2016**, *6*, 26269.

https://doi.org/10.1021/acsnano.5c03454
ACS Nano XXXX, XXX, XXX−XXX