Retrieval-Augmented Generation and LLM Agents for Biomimicry Design Solutions

Christopher Toukmaji[1*†], **Allison Tee**[2*†]
[1]University of California, Irvine
[2]Stanford University
ctoukmaj@uci.edu, ateecup@stanford.edu

Abstract

We present BIDARA, a **B**io-**I**nspired **D**esign **A**nd **R**esearch **A**ssistant, to address the complexity of biomimicry – the practice of designing modern-day engineering solutions inspired by biological phenomena. Large Language Models (LLMs) have been shown to act as sufficient general-purpose task solvers, but they often hallucinate and fail in regimes that require domain-specific and up-to-date knowledge. We integrate Retrieval-Augmented Generation (RAG) and Reasoning-and-Action agents to aid LLMs in avoiding hallucination and utilizing updated knowledge during generation of biomimetic design solutions. We find that incorporating RAG increases the feasibility of the design solutions in both prompting and agent settings, and we use these findings to guide our ongoing work. To the extent of our knowledge, this is the first work that integrates and evaluates Retrieval-Augmented Generation within LLM-generated biomimetic design solutions.

Introduction

Biomimicry practitioners face design barriers such as hardships trying to find plausible biological strategies for engineering problems, and the extensive time used identifying design solutions from these biological strategies (Nagel 2014; Chen et al. 2021b). As a result, we turn to LLMs to act as research assistants to streamline the literature review process and biological abstraction cycle.

Instruction-tuned LLMs have been demonstrating an increase in adoption as assistants. One such use case of LLMs is for researchers to formulate new ideas or automate the search of related work and relevant papers. In practice, however, there are several shortcomings when using LLMs as research assistants. Namely, general-purpose LLMs are not typically updated frequently, so the research works provided may be outdated. Moreover, LLMs may hallucinate, causing them to cite and reference non-existent work. To combat both of these issues present when using LLMs, we integrate *Retrieval-Augmented Generation* (RAG) with a general-purpose LLM. RAG is the practice of retrieving data outside an LLM, then prompting the LLM with the retrieved data as context under the pretext that the additional context will guide the model into a more accurate response.

Another solution to mitigate hallucination issues in LLMs is LLM agents. LLM agents wrap around LLMs and are prompted with a user query and instructions to select an action from a list of tools until the query has been answered. We experiment with both RAG and LLM agents; the class of LLM agents we use are Reasoning-and-Action agents (ReAct) (Yao et al. 2023), where a reasoning step and an action step are intertwined.

Our contributions are two-fold:

1. We evaluate the quality of LLM-generated biomimetic design solutions in settings with and without Retrieval-Augmented Generation, and in settings with and without ReAct Agents, to determine which setting to implement in BIDARA.
2. We present our directions for ongoing work further expanding the RAG tool suite. This is sparked by both the results of the quality of biomimetic design solutions, as well as feedback from domain experts.

Background and Related Work

Biomimicry With Natural Language Processing

The most similar work to ours is that of Zhu, Zhang, and Luo (2022). The work fine-tunes a base GPT-3 Davinci model with a dataset of 221 biomimicry examples gathered from the AskNature website, then the fine-tuned model is used to generate design concepts. We complement the limitations of the work by using a large academic knowledge base for retrieval, and we use a more robust and capable LLM.

In the broader NLP space, there is more work on knowledge extraction for biomimetic analogy linking. Shu (2010) utilizes keyword frequency and knowledge extraction from a knowledge base to identify biomimetic analogies to design questions. Chen et al. (2021a) presents an algorithm that can be applied to generate keywords to search for relevant biological information. Chen et al. (2021b) presents a knowledge extraction method to identify biomimetic analogies.

Retrieval-Augmented Generation and Agents

RAG has largely been a successful strategy to minimize hallucination with LLMs (Shuster et al. 2021; Mallen et al.

[*]Work done during an internship at NASA Glenn Research Center.

[†]These authors contributed equally.

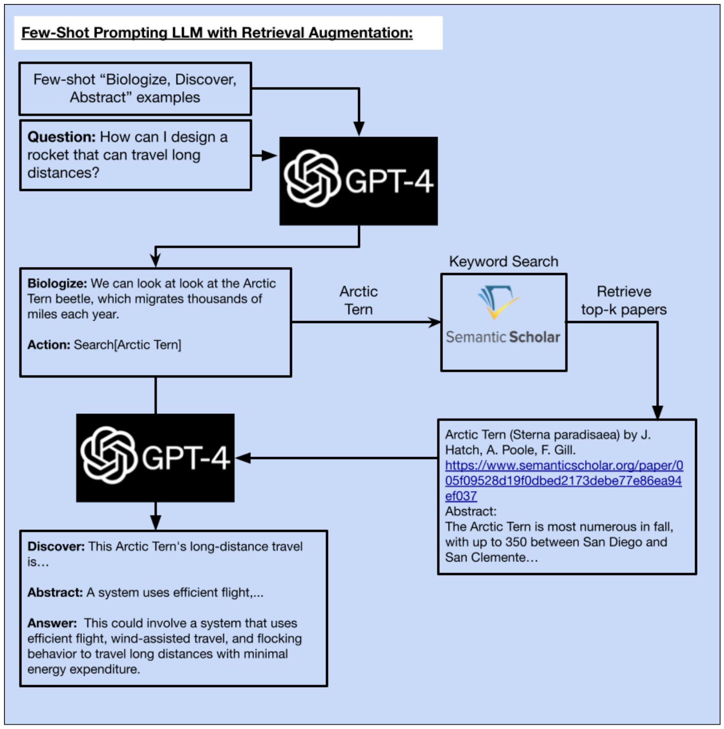


Figure 1: System Diagram of Few-Shot Prompting LLM + RAG.

2023; Komeili, Shuster, and Weston 2022). LLM agents, such as the ReAct Agents, wrap around pre-trained LLMs that do not use the internet. LLM agents have been applied in tasks where LLMs tend to hallucinate, such as arithmetic, source citation, or other tasks that would benefit from the use of the Internet or an external tool. Some use-cases of agents are for program assistance (Gao et al. 2023), multi-modal retrieval (Yang et al. 2023), and general-purpose API use (Liang et al. 2023).

Experimental Setup

We develop a test set of ten biomimicry design questions. Each question is evaluated in four experimental settings which are outlined below. The final solution is evaluated by human annotators in terms of feasibility and novelty. A system diagram of one of the experiment settings is given in Figure 1.

Experiments

We experiment in the following settings:

1. **Zero-Shot Prompting LLM:** The LLM is prompted with a user query appended to a prompt with the task description as a baseline.
2. **Few-Shot Prompting LLM + RAG:** The LLM is prompted twice in few-shot settings. The first prompt generates a search term which is queried to an academic paper database. The top-k papers are retrieved from the database, and the abstracts of the papers are appended to the prompt. The second pass conditions on the same context as the first pass, in addition to the retrieved paper abstracts, to generate the rest of the answer.
3. **ReAct Agent:** The LLM is given a list of descriptions of tools to use and decides which tools to use, following the zero-shot ReAcT framework (Yao et al. 2023).
4. **ReAct Agent + RAG:** The LLM is given the same list of tools as the ReAct Agent, but RAG is added as a tool.

Model

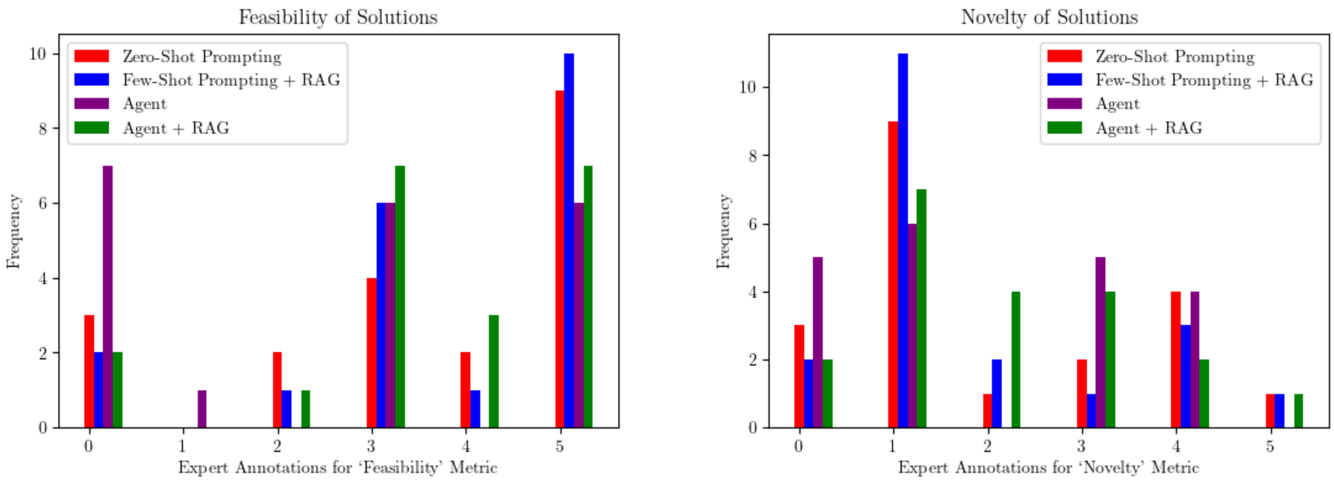
We use GPT-4 (OpenAI 2023b) with zero temperature as the base LLM for all experiments as it has been trained on the largest amount of data, so the model has stored a larger view of world understanding which would help with design solutions. Moreover, GPT-4 has a large number of parameters, so it is a robust model capable of zero-shot and few-shot learning (Brown et al. 2020).

Retrieval Augmentation Module

We use the Semantic Scholar Graph API (Kinney et al. 2023) with keyword search as the RAG module for retrieving papers. Our motivation behind the use of the Semantic Scholar Graph API is that the API is updated with new papers frequently, and it enables access to the largest academic knowledge base at the time of writing - with over 200 million papers available.

Metrics

Design solutions are evaluated by human evaluators on both feasibility and novelty on a scale of 0 to 5. We use a similar



a) Histogram of Feasibility Scores

b) Histogram of Novelty Scores

Figure 2: Experts' Annotations of generated design solutions for the design questions in the test set.

rubric as Zhu, Zhang, and Luo (2022) which spans from 1 to 5. A feasibility score of 1 denotes that the generated design solution makes no sense and is infeasible, while a feasibility score of 5 denotes that the generated design solution makes perfect sense and is completely feasible. Similarly, a novelty score of 1 denotes that the generated design solution already exists and is a common solution, while a novelty score of 5 denotes that the generated design solution is entirely novel.

We introduce a score of 0 in both feasibility and novelty metrics, which is reserved for design solutions that do not answer the design question. Our motivation behind the introduction of this score is that an unrelated solution should be scored lower than a related, but low-quality, solution.

Human Evaluation

Each annotator is presented with a biomimicry design question and four outputs, one for each of the prompting methods. We report the average score and standard deviation across the test dataset which contains ten biomimetic design questions. For scalable evaluation, we present the generated design answer and the generated explanation of the answer to the annotator, leaving out the chained intermediate steps. To avoid any bias towards preferring a design solution that refers to a paper, we omit all intermediate steps except for the final answer, and we manually remove any references to cited papers within the answer. Due to resource constraints, we have two biomimicry practitioners as human annotators.

Methods

We further describe the four settings used in experiments.

Zero-Shot Prompting LLM

The LLM is prompted to generate a design solution in zero-shot with a natural language description of the first three steps of the Biomimicry Design Toolbox (Biomimicry Institute 2023): (1) Biologize, (2) Discover, (3) Abstract.

Few-Shot Prompting LLM + RAG

The LLM is prompted to generate a design solution in a few-shot ReAct prompt with three examples of the following reasoning sequence: generating a search query, retrieving top-k abstracts, and using the first three steps of the Biomimicry Design Toolbox - (1) Biologize, (2) Discover, (3) Abstract. There are two LLM inference steps. The first pass prompts the LLM to generate the Biologize step and a search query to pass to the RAG module. The RAG module searches the Semantic Scholar API and retrieves the top-k papers. The titles, authors, URLs, and abstracts of each of the top-k papers are appended to the few-shot prompt and the output of the Biologize step. The LLM is then prompted again to complete the Discover and Abstract step. The prompt does not include any additional instructions. We find that setting $k = 2$ is the maximum value of k that consistently fits the paper abstracts within the context window.

ReAct Agent

We utilize a zero-shot ReAct framework to develop a Reasoning-and-Action Agent. The agent is given a natural language description of the steps from the Biomimicry Design Toolbox - (1) Biologize, (2) Discover, (3) Abstract, (4) Emulate, (5) Evaluate - as agent tools. When one of the tools is selected, a natural language description of the toolbox step is passed to the LLM. To prevent the agent chain from exceeding the context window, we pass the initial zero-shot prompt and the most recent reasoning step at every time step to the agent. The tools in this experiment are natural-language prompts and do not make use of the Internet.

ReAct Agent + RAG

We utilize an identical setup as the ReAct Agent setting, but RAG is added as a tool. When the RAG tool is selected, the RAG module searches the Semantic Scholar API with a

Question: How can we design a vehicle to travel from Sydney to USA in one hour?

Biologize: The essential function we need to address is rapid long-distance travel. In biological terms, we can ask, "How does nature travel long distances quickly?" Specifically, we can look at the Peregrine Falcon, the fastest bird in the world, which uses gravity to its advantage during its hunting stoop (high-speed dive). So, our biologized question becomes, "How does the Peregrine Falcon achieve high-speed travel?"

Action: Search[Peregrine Falcon high-speed travel]
Shape Optimization of High-Speed Rail by Biomimetic by C. T. Foo, B. Omar, I. Taib.
Abstract: The Southern Corridor High-Speed Rail (HSR) has the potential to connect Kuala Lumpur and Singapore with only 90 minutes of travel time. The HSR needs to travel at minimum drag in order to promote ride comfort and environmental sustainability. Eight combinations of train head bio-models are generated from a morphological chart of three key concepts formed by biomimetic method, referring to the peregrine falcon, belted kingfisher, sailfish and dolpine. The three key concepts are the type of train nose, cross-section shape, and train nose shape. The bio-models are drawn by computer-aided drawing software and the air flow around the bio-models are simulated in order to obtain the drag coefficients in counterflow condition. The drag coefficient of each bio-models are compared and validated with published drag coefficient of Harmony CRH380A, which is 0.1971. The bio-models with lowest drag coefficient is the bio-model with sharp-streamlined nose and round-based crosssection, which is 0.1559... (cntd.)
Russia's Peregrine Falcon Takes Flight by K. Barrow.
Abstract: Known in Russia as the Sapsan, or Peregrine Falcon, Russian Railway's (RZD's) first high-speed train is being launched this month with service between Moscow and St. Petersburg. Siemens was awarded a contract for a fleet of eight 10-car Sapsan trains in May 2006. Designed to operate in temperatures as low as -40 degrees Celsius, the Sapsan has a wider body shell than its western European counterparts, very strict criteria ....(cntd.)

Discover: The Peregrine Falcon achieves high-speed travel by using a streamlined body shape to reduce drag. This allows it to reach speeds of up to 240 miles per hour during its hunting stoop. Similarly, high-speed trains are designed with a streamlined shape to reduce air resistance and increase speed. The Russian Sapsan train, named after the Peregrine Falcon, is designed to operate in extreme temperatures and has a wider body shell than its western European counterparts to reduce drag.

Abstract: The essential feature that makes the Peregrine Falcon's high-speed travel successful is its streamlined body shape that reduces drag. In design terms, we can describe this strategy as follows: "A streamlined body shape is used to reduce air resistance and increase speed."

Answer: *A vehicle designed for rapid long-distance travel could mimic the streamlined body shape of the Peregrine Falcon to reduce drag and increase speed. This could involve a design with a sharp, pointed nose and a smooth, rounded body. This design could potentially improve the speed and efficiency of long-distance travel by reducing air resistance.*

Figure 3: Full output from the Few-Shot Prompting LLM + RAG experiment and the *answer presented to the annotators*. Human annotators labeled this response with a feasibility score of 5 and a novelty score of 1.

Method	Feasibility	Novelty
Zero-Shot Prompting	3.45 ± 1.77	1.90 ± 1.55
Few-Shot Prompting + RAG	3.70 ± 1.58	1.75 ± 1.41
ReAct Agent	2.45 ± 2.06	1.85 ± 1.53
ReAct Agent + RAG	3.50 ± 1.50	2.00 ± 1.34

Table 1: Experts' Annotations of generated design solutions for the design questions in the test set.

search query and retrieves the top-k papers. The titles, authors, URLs, and abstracts of each of the top-k papers are passed to the next reasoning step. We find that setting $k = 2$ is the maximum value of k that consistently fits the paper abstract and previous reasoning steps within the context window. To prevent the agent chain from exceeding the context window, we pass the initial zero-shot prompt and the most recent reasoning step at every time step to the agent. The RAG tool uses the Internet, but the other tools are natural-language prompts.

Results

Results from the expert annotations of all four experimental settings are displayed in Table 1 and Figure 2. There are several key takeaways after analysis of the distribution of the expert annotations. First, integrating RAG increases the feasibility of the design solution in both prompting and agent settings, but the novelty of the design solution varies. Moreover, methods with RAG are less likely to be unrelated to the design question. Lastly, prompting methods have higher feasibility and novelty scores in most cases than the agent counterparts. In turn, these results suggest that the Few-Shot Prompting LLM with RAG setting performs the best based on the experts' annotations and their feedback that the consideration of feasibility warrants a heightened emphasis. As a result, we opt to use the prompting LLM with RAG setting within BIDARA.

Directions and Motivation of Current Work

Function-Calling Methods

The approach for RAG and tool usage in BIDARA shifted to function-calling, which provides a suite of tools to aid the user in further exploration of their topic. Function-calling is a feature supported by the OpenAI API (OpenAI 2023a). BIDARA has access to five different functions: paper retrieval, setting/querying research space, patent search, and image generation. Paper retrieval, patent search, and image generation use a straightforward API call structure while the research space functions store and access information in LlamaIndex nodes (Liu 2022).

Frameworks

BIDARA with function-calling uses GPT-4 as the base LLM. In addition, we use a Google Search API to retrieve patents from Google Patents (SerpApi 2023). Images are generated by Dall-E 2, and the full text of papers is retrieved from Semantic Scholar and organized into a searchable database using LlamaIndex. The research space (paper Q&A) feature is based on code from S2QA (Rohatgi 2023), a project made in collaboration with Semantic Scholar.

Motivation and Testing

Biomimicry domain experts were invited to interact with BIDARA and feedback via user interviews was collected on their experiences with a preliminary few-shot version of BIDARA, such as liked and disliked features along with areas for improvement. This feedback was collected to allow experts to input challenges relevant to their work and steer the dialogue.

The accuracy and dependability of function-calling is tested during development. We ask targeted questions towards the research space, and paper Q&A outputs are manually compared to the information provided in the papers themselves to ensure accuracy.

Expert Feedback

One expert, a bio-inspired design researcher, and professor, states that she likes "the reminders about the process," the provided citations, and the ease of chatting. However, she notes that the response time is slower than other AI systems and that the suggested design strategies are limiting. She suggested that the citations should become hyperlinks to papers, a feature later implemented using RAG. The emphasis on fact-based generation is echoed by other experts. We address this concern with RAG, gathering research and patent papers to enable more in-depth responses.

Another interviewee, a consultant and university fellow for biomimicry, likes how BIDARA provides a detailed, structured response in the Biologize step and how it points out if a question is unfeasible and suggests rephrasing the question (e.g. "How can we design a vehicle to travel from Sydney to the U.S. in one hour?" → "How might we enable rapid, efficient, and sustainable transcontinental travel?"). However, she notices that sometimes, BIDARA misses important factors such as listing impacts and leverage points.

Other interviewees have mentioned that there are not enough visuals, a user experience issue that led to the addition of `generateImage()`. Although capable of general visualizations, BIDARA is unable to provide scientifically accurate design sketches, judged in comparison to expert diagrams in the same context (Biomimicry Institute 2023). For instance, when asking BIDARA to display an image of "a jackrabbit using its large ears to cool off," the assistant responds with a picture of a jackrabbit and a short description of blood vessel expansion in the ears dissipating heat. When asked to provide a diagram of the facilitation of heat loss, BIDARA outputs a vague depiction of branching vessels.

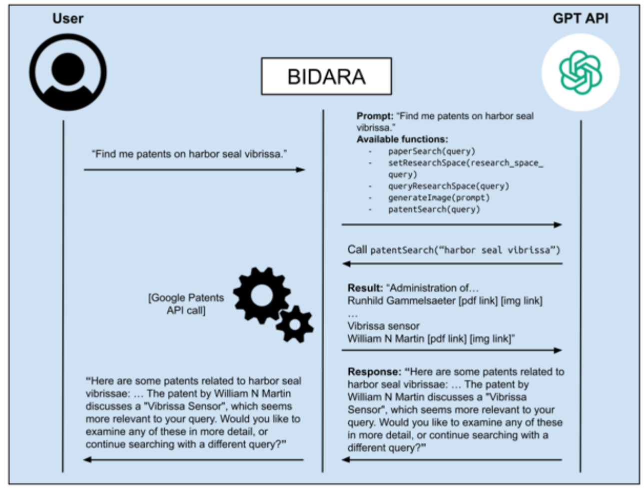


Figure 4: Function-calling retrieves patents in response to a user query on harbor seal whiskers.

Function-Calling Results

From tests performed during development, the function-calling framework calls the function implied by a query to a very high degree of accuracy. The paper Q&A responses also align with the papers in the research space. When requesting specific papers using ambiguous references ("According to Guo and Wang", "In Cui et. al.",...), specific information from the correct papers is also reliably retrieved.

Conclusion

The development of BIDARA marks an advancement in the field of AI tools for biomimicry. This study integrates ReAct agents and RAG to address the key issues of hallucination and outdated knowledge in LLMs, particularly when applied as research assistants in biomimetic design. Our results demonstrate that the inclusion of RAG enhances the feasibility of design solutions generated by LLMs in both prompting and agent settings. Notably, the Few-Shot Prompting LLM + RAG emerged as the most feasible method, and Agent + RAG was the most novel. Moreover, the ongoing incorporation of a more diverse RAG toolkit provides an expanded, interactive capability. This work not only introduces a formal study of biomimicry via LLMs but also furthers the integration of more sophisticated AI techniques in the field.

Acknowledgements

This work was conducted as part of the Periodic Table of Life (PeTaL) project at the NASA Glenn Research Center (Shyam et al. 2019). We extend our heartfelt gratitude to our mentors, Brandon Ruffridge and Herbert Schilling, for their guidance, expertise, and support throughout our work on this project. Finally, we are grateful to the many biomimicry experts for sharing their knowledge and providing critical feedback on BIDARA.

References

Biomimicry Institute. 2023. Abstract Design Strategies. https://toolbox.biomimicry.org/methods/abstract/. Accessed: 2024-02-26.

Biomimicry Institute. 2023. The Biomimicry Process - Biomimicry Toolbox — toolbox.biomimicry.org. https://toolbox.biomimicry.org/methods/process/. [Accessed 18-09-2023].

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Chen, C.; Li, Y.; Tao, Y.; Chen, J.; Liu, Q.; and Li, S. 2021a. A method to automatically push keywords for biological information searching in bio-inspired design. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 235(1): 30–47.

Chen, C.; Tao, Y.; Li, Y.; Liu, Q.; Li, S.; and Tang, Z. 2021b. A structure-function knowledge extraction method for bio-inspired design. *Computers in Industry*, 127: 103402.

Gao, L.; Madaan, A.; Zhou, S.; Alon, U.; Liu, P.; Yang, Y.; Callan, J.; and Neubig, G. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, 10764–10799. PMLR.

Kinney, R. M.; Anastasiades, C.; Authur, R.; Beltagy, I.; Bragg, J.; Buraczynski, A.; Cachola, I.; Candra, S.; Chandrasekhar, Y.; Cohan, A.; Crawford, M.; Downey, D.; Dunkelberger, J.; Etzioni, O.; Evans, R.; Feldman, S.; Gorney, J.; Graham, D. W.; Hu, F.; Huff, R.; King, D.; Kohlmeier, S.; Kuehl, B.; Langan, M.; Lin, D.; Liu, H.; Lo, K.; Lochner, J.; MacMillan, K.; Murray, T.; Newell, C.; Rao, S. R.; Rohatgi, S.; Sayre, P. L.; Shen, Z.; Singh, A.; Soldaini, L.; Subramanian, S.; Tanaka, A.; Wade, A. D.; Wagner, L. M.; Wang, L. L.; Wilhelm, C.; Wu, C.; Yang, J.; Zamarron, A.; van Zuylen, M.; and Weld, D. S. 2023. The Semantic Scholar Open Data Platform. *ArXiv*, abs/2301.10140.

Komeili, M.; Shuster, K.; and Weston, J. 2022. Internet-Augmented Dialogue Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8460–8478. Dublin, Ireland: Association for Computational Linguistics.

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

Liu, J. 2022. LlamaIndex. https://github.com/jerryjliu/llama_index. Accessed: 2024-02-26.

Mallen, A.; Asai, A.; Zhong, V.; Das, R.; Khashabi, D.; and Hajishirzi, H. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9802–9822. Toronto, Canada: Association for Computational Linguistics.

Nagel, J. K. S. 2014. *A Thesaurus for Bioinspired Engineering Design*, 63–94. London: Springer London. ISBN 978-1-4471-5248-4.

OpenAI. 2023a. Function Calling - OpenAI Platform Documentation. https://platform.openai.com/docs/guides/function-calling. Accessed: 2024-02-26.

OpenAI. 2023b. GPT-4 Technical Report. arXiv:2303.08774.

Rohatgi, S. 2023. S2QA: Question Answering on research papers from Semantic Scholar.

SerpApi. 2023. Google Search Results in Python. https://github.com/serpapi/google-search-results-python. Accessed: 2024-02-26.

Shu, L. 2010. A Natural Language Approach to Biomimetic Design. *AI EDAM*, 24: 507–519.

Shuster, K.; Poff, S.; Chen, M.; Kiela, D.; and Weston, J. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3784–3803. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Shyam, V.; Friend, L.; Whiteaker, B.; Bense, N.; Dowdall, J.; Boktor, B.; Johny, M.; Reyes, I.; Naser, A.; Sakhamuri, N.; Kravets, V.; Calvin, A.; Gabus, K.; Goodman, D.; Schilling, H.; Robinson, C.; Reid II, R. O.; and Unsworth, C. 2019. PeTaL (Periodic Table of Life) and Physiomimetics. *Designs*, 3(3).

Yang, Z.; Li, L.; Wang, J.; Lin, K.; Azarnasab, E.; Ahmed, F.; Liu, Z.; Liu, C.; Zeng, M.; and Wang, L. 2023. MM-REACT: Prompting ChatGPT for Multimodal Reasoning and Action. arXiv:2303.11381.

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

Zhu, Q.; Zhang, X.; and Luo, J. 2022. Biologically Inspired Design Concept Generation Using Generative Pre-Trained Transformers. arXiv:2212.13196.