VISUAL AI AND LINGUISTIC INTELLIGENCE THROUGH STEERABILITY AND COMPOSABILITY

David Noever and Samantha Elizabeth Miller Noever

PeopleTec, 4901-D Corporate Drive, Huntsville, AL, USA, 35805

ABSTRACT

This study explores the capabilities of multimodal large language models (LLMs) in handling challenging multistep tasks that integrate language and vision, focusing on model steerability, composability, and the application of long-term memory and context understanding. The problem addressed is the LLM's ability (Nov 2023 GPT-4 Vision Preview) to manage tasks that require synthesizing visual and textual information, especially where stepwise instructions and sequential logic are paramount. The research presents a series of 14 creatively and constructively diverse tasks, ranging from AI Lego Designing to AI Satellite Image Analysis, designed to test the limits of current LLMs in contexts that previously proved difficult without extensive memory and contextual understanding. Key findings from evaluating 800 guided dialogs include notable disparities in task completion difficulty. For instance, 'Image to Ingredient AI Bartender' (Low difficulty) contrasted sharply with 'AI Game Self-Player' (High difficulty), highlighting the LLM's varying proficiency in processing complex visual data and generating coherent instructions. Tasks such as 'AI Genetic Programmer' and 'AI Negotiator' showed high completion difficulty, emphasizing challenges in maintaining context over multiple steps. The results underscore the importance of developing LLMs that combine long-term memory and contextual awareness to mimic human-like thought processes in complex problem-solving scenarios.

KEYWORDS

Large language models, creativity, steerability, composability, dataset

1. INTRODUCTION

The creative maker spaces have become vibrant hubs of 21st-century innovation, merging the traditional tactile experience with digital fabrication and design. However, integrating new artificial intelligence (AI) tools and, in particular, the current generation of multimodal large language models (LLMs)[1-17] into these environments has the potential to enhance human creativity and innovation [18-32]. In recent years, the intersection of AI and multimodal (MM) learning has spawned a generation of models that integrate and interpret information across various forms of data, including text, images, and speech. In short, AI models now combine both vision and language understanding [1]. These models promise new approaches for human-computer interaction, complex problem-solving, and decision-making processes. Along with competitors like Google Bard, Open AI's GPT-4 Vision development lays a critical foundation, enhanced language understanding and vision models that generate and analyze imagery [1]. Table 1 summarizes the current Open AI challenge list of multimodal LLM shortcomings [1] yet to be mastered by current models.

David C. Wyld et al. (Eds): CNSA, DMDB, NIAI, ICCIoT, AIAP -2024
pp. 61-72, 2024. CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.140104

Table 1. Areas for Multimodal Large Language Models to Evolve New Capabilities

Data Challenge	Evaluation
Medical images	The model is unsuitable for interpreting specialized medical images like CT scans and shouldn't be used for medical advice.
Non-English	The model may not perform optimally when handling images with text of non-Latin alphabets, such as Japanese or Korean.
Big text	Enlarge text within the image to improve readability, but avoid cropping essential details.
Rotation	The model may misinterpret rotated / upside-down text or images.
Visual elements	The model may struggle to understand graphs or text where colors or styles like solid, dashed, or dotted lines vary.
Spatial reasoning	The model struggles with tasks requiring precise spatial localization, such as identifying chess positions.
Image shape	The model struggles with panoramic and fisheye images.
Metadata and resizing	The model doesn't process original file names or metadata; images are resized before analysis, affecting their dimensions.
Counting	It may give approximate counts for objects in images.
CAPTCHA S	For safety reasons, we have implemented a system to block the submission of CAPTCHAs.

Despite the potential symbiosis between multimodal LLMs and creative maker spaces, there remains a discernible gap in the seamless integration of these advanced AI systems into the iterative, hands-on environments that characterize maker spaces. As inventor Thomas Edison remarked anecdotally, "*I haven't failed. I've just found ten thousand different ways that don't work.*" While robust in knowledge and pattern recognition, current multimodal LLMs often fall short in their capacity for intuitive generative design and adaptability to the highly variable context of maker spaces.

Soon after ChatGPT's release [1], Psychology Today [33] advised teachers to recraft curriculum away from basic knowledge tests and further embrace creative tasks that might challenge the current AI generators: "*From an instructional perspective, in addition to using AI detection software, focus on assessments that evaluate creativity or apply knowledge in specific contexts instead of testing for accuracy alone. Avoid the use of knowledge recognition and recall through the elimination of multiple-choice questions.*"

However, the commonsense approach to scoring AI performance as "mechanical thinking" has not yet borne itself out. As Open AI's Sam Altman [34] succinctly summarized: "*Creativity has been easier for AI than people thought.*"

One can recast the tension between AI steerability ("it follows my instructions") and composability ("assembly by related combinations") in these tests, with a hypothesis that in its loosest form, unexplainable machine intelligence as expressible creativity may ultimately prove as unexplainable and bewildering as human creativity itself.

To explore these somewhat unanticipated functional creative gains, we conducted a series of experimental case studies to explore the extent to which these LLMs can engage in co-creative tasks, learn from iterative design cycles, and contribute to innovation within the diverse ecosystem of a maker space. The primary aim of this study is to systematically analyze the performance of vision integrated LLMs in various creative contexts. By doing so, we seek to extend Table 1 with their strengths, weaknesses, and potential areas for improvement within the collaborative creativity and maker culture framework. A critical experimental goal highlights both the steerability ("model follows instructions") and composability ("model rearranges according to instruction").

To effectively solve the creative multi-step challenge of LLM memory coupled with the image-language boundary, our research initiative focuses on testing and assessing an advanced multimodal LLM like GPT4-vision preview (November 2023). This research challenge would integrate the diverse capabilities demonstrated across various professional fields, navigate complex relational reasoning within visual contexts, and secure multimodal communication. It addresses the critical issue of enhancing LLM memory for sequential, multi-step tasks while bridging the gap between image and language processing, setting a repeatable benchmark [32] in AI-driven innovation and problem-solving.

2. METHODS

The research explores multimodal LLMs with applications that center around building or designing novel outcomes from a prescribed set of initial elements. We focus those elements on product design, culinary arts, and educational sectors. We prompt the vision integrated LLM to perform multiple creative tasks borrowed from the spirit or curriculum of maker spaces. We selected the tasks to maximize the diversity and range of creative and technical skills represented but filtered by outcomes that the LLM could deliver as a novel series of generated texts or images.

2.1. Approach

As released in November 2023 the vision-enabled GPT-4 model allows researchers to process visual input and generate relevant outputs for different creative tasks. Inspired by maker movements, we equip a set of tasks with the contextual prompts needed for toolmaking, material design, cooking, Lego building (Figure 1), and others. Appendices highlight the pre- and post-development of the job, mainly as qualitative use cases and examples. Where appropriate, the latter stages of each task generate an interaction log, an image, a recipe, or a miniature movie or animation gif.

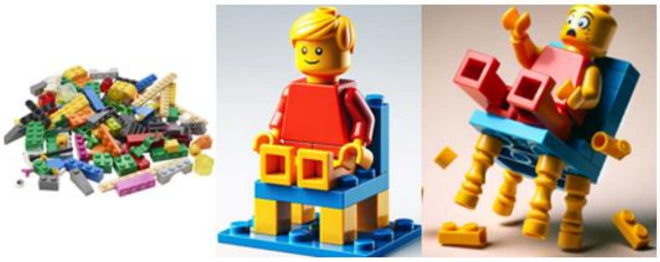

Figure 1. Logic Progression from Raw Inputs to Steerable Conception, LEGO Example

A formative baseline for designing assignments might include a timeline, whether the model gets just an image of a refrigerator or restaurant bar, catalogs the ingredients using its built-in object detection, and then generalizes under user instruction and interaction to refine the final product, which represents a meal plan, mixology instruction or general construction advice. Table 2 summarizes the tasks and their estimated difficulty and expected outputs. The problem is a subjective evaluation based on the number of required reasoning steps, memory, and complexity of the result. The assessment of a successful outcome follows from interaction logs (Appendices), which catalog the number of successful task completions, user corrections, and time (steps) taken to bridge intermediate design stages. The analysis emphasizes a thematic study to discover categories of tasks that Vision-Language struggles with currently and, where appropriate, assesses the efficiency of task completion via the number of instruction steps needed to generate the expected outputs.

3. Results

The Appendices A-N [40] show the full dialog and LLM response from Table 2 vignettes and challenge problems. Of the 14 major tasks, the dialogues cover 800 sequences of human-machine instruction and collaborative assembly and disassembly of either physical or virtual objects. To test steerability, scoring successes was simplified to include whether the task was completed as defined through iterative responses. To test composability, scoring accomplishments center on whether the LLM disassembled or assembled parts to complete a picture, text challenge, or ingredient recipe.

The curated tasks outlined in Table 2 for stress-testing multimodal large language models (LLMs) exhibit diverse challenges, showcasing the model's proficiency in both language and vision-based studies. These tasks are grouped into two main categories: 'Image to Instruction' and 'Stepwise Text to Text/Image,' each requiring a unique blend of creative and analytical skills.

Table 2. Curated Tasks for Stress-Testing Multimodal Large Language Models

Appendix: Task	Creative Skill Tests	Predicted Completion Difficulty
A: AI Lego Designer	Image to Instruction	MID
B: AI Aerospace Designer	Image to Instruction	MID
C: Image to Ingredient AI Bartender	Image to Instruction	LO
D: Image to Ingredient AI Chef	Image to Instruction	MID
E: Image to AI Origami (Japanese Art of Folding)	Image to Instruction	MID
F: Image to AI Kintsugi (Art of Repairing Broken Objects)	Image to Instruction	MID
G: AI Negotiator	Stepwise Text to Text	HI
H: AI Cyber Defender	Stepwise Text to Text	MID
I: AI Three-Panel Cartoonist	Stepwise Text to Text	MID
J: AI Genetic Programmer	Stepwise Text to Text	HI
K: AI Excel Spreadsheet Chartist	Stepwise Text to Image	LO
L: AI Salad Chef	Image to Instruction	MID
M: AI Game Self-Player	Stepwise Text to Image	HI
N: AI Satellite Image Analyst	Image to Instruction	HI

Groupings and Highlights

1. Image to Instruction Tasks

- **AI Lego and Aerospace Designers (Mid Difficulty):** These tasks involve converting images into structured instructions, demonstrating the model's ability to interpret complex visual data and translate it into coherent, step-by-step guidance.
- **AI Bartender and Chef (Low to Mid Difficulty):** These tasks focus on translating images into ingredients, showcasing the model's ability to analyze visual data and extract relevant information.

- **AI Origami, Kintsugi, Pantry Chef, and Satellite Image Analyst (Mid to High Difficulty):** These tasks require a deeper understanding of cultural contexts (Origami, Kintsugi) and specialized knowledge (satellite imagery), pushing the model's capabilities in processing and instructing based on images.

2. Stepwise Text to Text/Image Tasks

- **AI Negotiator and Genetic Programmer (High Difficulty):** These tasks demand advanced logical reasoning and complex problem-solving skills in a stepwise text format.
- **AI Cyber Defender, Three-Panel Cartoonist, and Excel Spreadsheet Chartist (Mid to Low Difficulty):** These tasks, ranging from cybersecurity to creating visual content, test the model's ability to process and sequentially generate text and images.
- **AI Game Self-Player (High Difficulty):** This unique task requires the model to interact with a game environment, demonstrating its potential in dynamic decision-making and strategy development.

4. DISCUSSION AND RELATION TO PREVIOUS LITERATURE

A milestone towards Artificial General Intelligence (AGI) involves a machine model that reasons over many diverse human tasks, not just specializations like playing chess, strategizing in Go, or solving puzzles. The present contributions intentionally support various ranges of challenging multimodal tasks to explore the steerability and composability of this LLM generation. The vignettes collectively reveal the breadth of multimodal LLMs' potential applications, from job-specific training to advanced reasoning in question answering and even the visual interpretation and synthesis of complex data (Figures 2-3). A multimodal LLM that stands at the vanguard of this evolving landscape would not only need to integrate these diverse capabilities but also navigate the pitfalls identified. It would need to be adept at specialized tasks across professions [13], capable of relational reasoning within visual contexts [14], and proficient in interpreting and generating multimodal scientific communications [15]. Additionally, it must be fortified against adversarial vulnerabilities [16], suggesting a development path that prioritizes robustness and contextual sensitivity as much as intellectual agility.

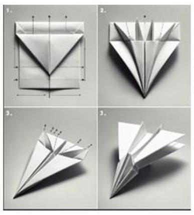

Figure 2. Fold Instruction Progression to Steerable Conception, Paper Airplane Example

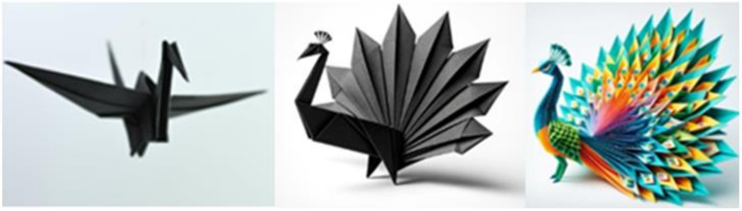

Figure 3. Progression from Decomposition to Recompositing

The notable features of this work, particularly in the context of challenging large language models (LLMs) with distinctly human tasks, revolve around steerability and composability. These tasks, exemplified by the LEGO analogy, demonstrate the LLM's ability to navigate between assembling and deconstructing components, whether images or text instructions. This approach tests the LLM's technical capabilities and probes its creative faculties [40].

Steerability and Composability in LLMs

- **Steerability refers to the model's ability to be guided or directed toward a specific outcome or** follow a set of instructions. The model must follow and generate structured instructions based on visual inputs in tasks like the AI Lego Designer (Appendix A) or AI Aerospace Designer (Appendix B).
- **Composability:** This involves the model's ability to combine disparate elements to create something new. In tasks like the AI Origami (Appendix E) or AI Kintsugi (Appendix F), the model must piece together information from images to create comprehensive instructions, showcasing digital creativity.

Creativity in LLMs: The LEGO Analogy

The LEGO analogy (Figure 1, Appendix A) is an apt metaphor for the creative process in LLMs. Just as LEGOs are assembled from individual blocks into a cohesive whole, LLMs piece together disparate elements of data (text, images) to generate new, coherent outcomes (Figure 4, Appendix F). This process reflects a form akin to human creativity, where new ideas often emerge from combining and recombining existing ones.

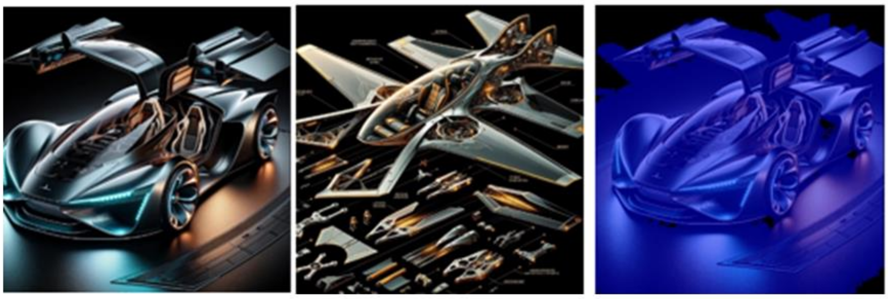

Figure 4. Progression from One Conception to a Revised Version and Difference

4.1. Maker Spaces and Automated Creative Innovation

Realizing a new paradigm of creation and learning blurs the human-machine creative boundaries of creativity. Gong et al. extend a multi-sensory approach by presenting multimodal-GPT, a model combining vision and language for dialogue, stepping towards more natural human-computer communication [2]. Kurti et al. [17] emphasize the importance of practical implementation in educational maker spaces, which intelligent, contextually aware systems might enhance. Keone and Peppler [18] reflect on the materials and tools that define a maker space, which AI could dynamically optimize to inspire and facilitate new forms of making. Knibbe et al. [19] propose the concept of a Smart Makerspace, an immersive instructional space for physical tasks, in which multimodal LLMs might provide real-time, adaptive instruction and feedback. Browder et al. [20] explore product development within corporate maker spaces, a frontier where multimodal LLMs can foster hybrid innovation logic by bringing together diverse knowledge bases and perspectives. This collective research underscores the transformative potential of multimodal LLMs in reshaping the landscape of creativity, learning, and innovation across various domains.

4.2. AI Spaces for Growth Mindsets

These diverse applications and explorations collectively suggest a future where multimodal LLMs augment human creativity and learning and become integral components in the continuous evolution of maker spaces and creative industries. As Steier and Young [21] discussed, growth mindset theories align with maker spaces' iterative, experimental nature. This mentality resonates with the ethos of The Invention Studio described by Forest et al. [22], where the maker culture thrives on the freedom to experiment and learn through trial and error—an approach that AI companions could catalyze while learning alongside humans. Bubeck et al. [23] present early experiments with GPT-4, providing sparks of what they describe as nascent artificial general intelligence. These experiments demonstrate AI's potential to learn and apply knowledge and ideate and innovate, akin to human creativity. This notion parallels the work of Busov et al. [24] and Koza et al. [25], who explored systematic methods of engineering creativity through TRIZ and genetic programming, respectively. The LEGO/logo project by Resnick and Ocko showcases learning through design [26], a concept that AI capable of generating and iterating on design patterns could expand. Similarly, Koszewska and Bielecki's work [27] on component standardization in furniture design and Morris et al.'s review[28] of origami-inspired products illustrate the rich potential for multimodal LLMs to assist in the development of sustainable and innovative design practices.

In education, Frydenberg et al. demonstrate the value of teaching agile methodologies through paper airplanes [29], a learning experience in which LLMs could assist in understanding the principles of flight and design in real time. Moreover, Cromwell et al.'s [30] exploration of computational creativity in the culinary arts signifies a domain ripe for introducing AI that can analyze recipes and contribute to creating new culinary experiences (Figure 5, Appendix C-D, L).

The systematic review of lean simulation games in construction by Bhatnagar et al. [31] underscores the potential for LLMs in training and simulation, enhancing learning experiences through interactive and adaptive challenges. Finally, Gadre et al.'s [32] search for next-generation multimodal datasets highlights the need for comprehensive, rich data sources that can fuel the creative engines of these LLMs, enabling them to understand and contribute to the maker space ecosystem in meaningful ways.

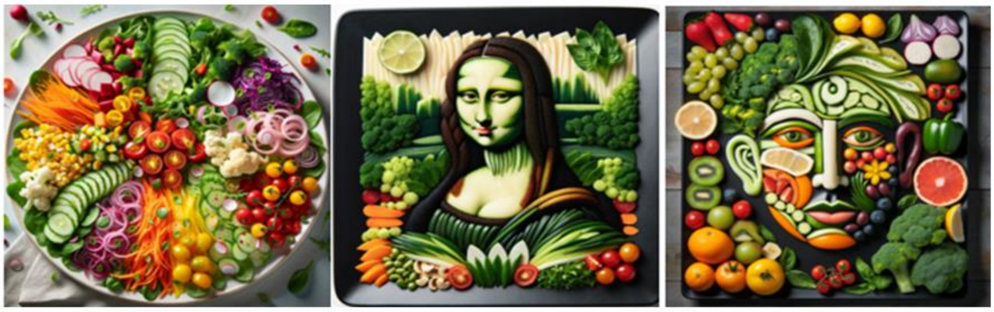

Figure 5. Progression of Ingredients to Style Transfer, AI Salad Chef example

4.3. Novel Roles for Multimodal Large Language and Vision Models

Bridging the gap between human ingenuity and AI capabilities in various fields necessitates a detailed examination of the current state and future potential of multimodal large language models (LLMs). AI systems often lack the nuanced understanding of materials, tools, and user intentions that are second nature to humans. AI researchers highlight a stark gap in developing collective intelligence systems that augment human ingenuity. There is a pressing need for vision-large language models that are both contextually aware and capable of co-creative processes, adapting and learning in tandem with their human counterparts within the dynamic, often unstructured confines of maker spaces. Such advancements would enhance creative collaboration and accelerate the innovation cycle, ultimately leading to a fusion of human and machine-driven innovation.

The culinary arts have also benefited [3], illustrating how AI can generate complex recipes from imagery, pushing the boundaries of creative AI applications [3]. Subsequent work dives deeper into visual reasoning, comparing the capabilities of Google Bard and GPT-Vision, which underscores the necessity for sophisticated multimodal analyses [4]. Wu et al. introduce Next-GPT, a model that transcends the limitations of modality by facilitating any-to-any conversion among multimodal inputs [5].

The medical field represents many AI-enabled government inventions. Wu et al. examine GPT-4V's potential in multimodal medical diagnosis, demonstrating substantial promise yet revealing limitations in high-stakes domains [6]. Echoing this view, Yang et al. assessed the performance of Multimodal GPT-4V in medical licensing exams, particularly in imaging diagnostics, offering a glimpse into future support systems for medical professionals [7]. The concept of "Socratic models" by Zeng et al. brings forth the idea of zero-shot multimodal reasoning, allowing models to compose answers from disparate sources without explicit training [8].

However, despite these advances, gaps remain. The accuracy of these models in diagnostic scenarios is a crucial concern, as demonstrated by Sorin et al., who focus on the diagnostic precision of GPT's multimodal analysis, suggesting that while there is potential, the path forward demands rigorous validation [11]. Finally, Yang et al. introduce Idea2Img, a self-refining approach using GPT-4V for iterative image design, which could potentially fill the creative gaps in automatic content generation [12].

In scoring LLM progress researchers have taken a pragmatic approach, curating a benchmark dataset aimed at professional certification, which could be pivotal in training LLMs for specialized job functions [13]. This push towards practical applications is complemented by Cadene et al.'s MUREL, which exploits multimodal relational reasoning, illuminating the path toward more contextually aware visual question-answering systems [14]. But as the original advice to teachers and curriculum designers pointed out, such tests involving multiple choices represent easy wins for LLMs that have previously seen vast quantities of training data and understand basic instructions [33].

Fernández-Fontecha et al. contribute a different perspective by examining visual thinking through a multimodal lens, emphasizing the role of visual notes in scientific communication [15]. Such qualitative enrichments to the data these models process could be vital in bridging the gap between human cognitive methods and AI processing. Moreover, the vulnerabilities of multimodal systems are subject to customized adversarial attacks on multimodal neurons, a reminder of the robustness yet to be achieved in these systems [16]. These studies illustrate the rapidly evolving landscape of multimodal LLMs, highlighting their growing impact across diverse domains, from culinary arts to medical diagnostics, while emphasizing the need for continuous improvement and robustness in these intelligent systems.

4.4. Scientifically Testable Hypotheses on Human Creativity

Some elements of human creativity seem to involve recombining and mashing diverse elements to produce a novel and pleasing outcome. This process resembles cognitive LEGO building, where the brain combines disparate ideas, concepts, or experiences to form new creations. To formalize this in a methodology that LLMs can interpret, we propose future efforts to enlarge the testability initiatives shown here.

Testability

1. **Experimental Design:** Conduct experiments where both human and machine participants are given sets of unrelated elements (images, words, concepts) and asked to create a cohesive story, artwork, or product.
2. **Creativity Measurement:** Use standardized creativity assessment tools to evaluate the outcomes based on originality, complexity, and aesthetic or functional value [35-39].

If the hypothesis holds, we expect to see a correlation between the ability to combine disparate elements and higher scores on creativity assessments effectively. Additionally, if validated, this hypothesis could provide insights into the mechanics of creativity in humans and AI and help refine AI models to mimic human creative processes better.

5. CONCLUSIONS

In conclusion, while large language models (LLMs) like GPT-4 Vision or Google Bard Vision offer remarkable capabilities in processing and interpreting a wide range of data, specific tasks present significant challenges (Table 1). For example, these models are admittedly unsuited for interpreting specialized medical images, such as CT scans, and should not be relied upon for medical advice. Their performance is less optimal with non-Latin alphabets, and there are difficulties in handling images with enlarged text or complex visual elements, like varying colors or line styles in graphs. The model also struggles with rotated or upside-down text and pictures and shows limitations in tasks requiring precise spatial reasoning, such as chess position identification. Panoramic and fisheye images pose a challenge due to their unique shapes.

Additionally, the inability to process metadata and the resizing of images before analysis can impact the interpretation of the original dimensions. The model provides only approximate counts for objects in pictures and, for safety reasons, cannot process CAPTCHAS. These limitations highlight the need for careful consideration when employing LLMs in specific contexts and underscore the importance of human oversight in scenarios where precision and specialized knowledge are critical.

Future tasks could integrate more complex combinations of these skills. For instance, tasks that blend the 'Image to Instruction' format with 'Stepwise Text to Text' challenges could be proposed. An example might be an AI Architect, where the model must interpret architectural designs (images) and provide stepwise construction guidelines (text). Another intriguing area could be tasks that require simultaneous image and text processing, such as AI Art Critic, where the model analyzes artwork (photo) and provides a detailed critique or historical context (text).

Moreover, increasing the complexity within each category, such as AI Medical Diagnostician, which would require interpreting medical images and providing stepwise medical advice, could significantly test the model's capabilities. These tasks challenge the model's current abilities and pave the way for exploring the limits of AI in creative and analytical domains, potentially leading to innovative applications in various fields.

ACKNOWLEDGMENTS

The authors thank the PeopleTec Technical Fellows program for encouragement and project assistance.

REFERENCES

[1] OpenAI, R. (2023). GPT-4 technical report. arXiv 2303.08774.

[2] Gong, T., Lyu, C., Zhang, S., Wang, Y., Zheng, M., Zhao, Q., ... & Chen, K. (2023). Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*.

[3] Noever, D., & Noever, S. E. M. (2023). The Multimodal and Modular AI Chef: Complex Recipe Generation From Imagery. *arXiv preprint arXiv:2304.02016.*

[4] Noever, D. A., & Noever, S. E. M. (2023). Multimodal Analysis of Google Bard And GPT-Vision: Experiments In Visual Reasoning. *arXiv preprint arXiv:2309.16705v2*

[5] Wu, S., Fei, H., Qu, L., Ji, W., & Chua, T. S. (2023). Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

[6] Wu, C., Lei, J., Zheng, Q., Zhao, W., Lin, W., Zhang, X., ... & Xie, W. (2023). Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. *arXiv preprint arXiv:2310.09909*.

[7] Yang, Z., Yao, Z., Tasmin, M., Vashisht, P., Jang, W. S., Wang, B., ... & Yu, H. (2023). Performance of Multimodal GPT-4V on USMLE with Image: Potential for Imaging Diagnostic Support with Explanations. *medRxiv*, 2023-10.

[8] Zeng, A., Attarian, M., Ichter, B., Choromanski, K., Wong, A., Welker, S., ... & Florence, P. (2022). Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*.

[9] Yang, Z., Li, L., Lin, K., Wang, J., Lin, C. C., Liu, Z., & Wang, L. (2023). The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, *9*.

[10] Li, Y., Liu, Y., Wang, Z., Liang, X., Liu, L., Wang, L., ... & Zhou, L. (2023). A Comprehensive Study of GPT-4V's Multimodal Capabilities in Medical Imaging. *medRxiv,* 2023-11.

[11] Sorin, V., Glicksberg, B. S., Barash, Y., Konen, E., Nadkarni, G., & Klang, E. (2023). Diagnostic Accuracy of GPT Multimodal Analysis on USMLE Questions Including Text and Visuals. *medRxiv*, 2023-10.

[12] Yang, Z., Wang, J., Li, L., Lin, K., Lin, C. C., Liu, Z., & Wang, L. (2023). Idea2Img: Iterative Self-Refinement with GPT-4V (ision) for Automatic Image Design and Generation. *arXiv preprint arXiv:2310.08541*.

[13] Noever, D., & Ciolino, M. (2023). Professional Certification Benchmark Dataset: The First 500 Jobs For Large Language Models. *arXiv preprint arXiv:2305.05377.*

[14] Cadene, R., Ben-Younes, H., Cord, M., & Thome, N. (2019). Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1989-1998).

[15] Fernández-Fontecha, A., O'Halloran, K. L., Tan, S., & Wignell, P. (2019). A multimodal approach to visual thinking: the scientific sketchnote. *Visual Communication*, *18*(1), 5-29.

[16] Noever, D. A., & Noever, S. E. M. (2021). Reading Isn't Believing: Adversarial Attacks On Multimodal Neurons. *arXiv preprint arXiv:2103.10480*.

[17] Kurti, R. S., Kurti, D., & Fleming, L. (2014). Practical implementation of an educational makerspace. *Teacher Librarian*, *42*(2), 20.

[18] Keune, A., & Peppler, K. (2019). Materials-to-develop-with: The making of a makerspace. *British journal of educational technology*, *50*(1), 280-293.

[19] Knibbe, J., Grossman, T., & Fitzmaurice, G. (2015, November). Smart makerspace: An immersive instructional space for physical tasks. In *Proceedings of the 2015 International Conference on Interactive Tabletops & Surfaces* (pp. 83-92).

[20] Browder, R. E., Crider, C. J., & Garrett, R. P. (2023). Hybrid innovation logics: Exploratory product development with users in a corporate makerspace. *Journal of Product Innovation Management*, *40*(4), 451-474.

[21] Steier, L., & Young, A. W. (2016). Growth mindset and the makerspace educational environment.

[22] Forest, C. R., Moore, R. A., Jariwala, A. S., Fasse, B. B., Linsey, J., Newstetter, W., ... & Quintero, C. (2014). The Invention Studio: A University Maker Space and Culture. *Advances in Engineering Education*, *4*(2), n2.

[23] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

[24] Busov, B., Mann, D. L., & Jirman, P. (1999, May). TRIZ and invention machine: Methods and systems for creative engineering and education in the 21st century. In *proc. of the 1st International Conference on Advanced Engineering Design Methods, Prague*.

[25] Koza, J. R., Bennett III, F. H., & Stiffelman, O. (1999, May). Genetic programming as a Darwinian invention machine. In *European Conference on Genetic Programming* (pp. 93-108). Berlin, Heidelberg: Springer Berlin Heidelberg.

[26] Resnick, M., & Ocko, S. (1990). *LEGO/logo--learning through and about design*. Cambridge: Epistemology and Learning Group, MIT Media Laboratory.

[27] Koszewska, M., & Bielecki, M. (2020). How to make furniture industry more circular? The role of component standardisation in ready-to-assemble furniture. *Entrepreneurship and Sustainability Issues*, *7*(3), 1688.

[28] Morris, E., McAdams, D. A., & Malak, R. (2016, August). The state of the art of origami-inspired products: A review. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* (Vol. 50169, p. V05BT07A014). American Society of Mechanical Engineers.

[29] Frydenberg, M., Yates, D., & Kukesh, J. (2018). Sprint, then fly: Teaching agile methodologies with paper airplanes. *Information Systems Education Journal*, *16*(5), 22.

[30] Cromwell, E., Galeota-Sprung, J., & Ramanujan, R. (2015, April). Computational creativity in the culinary arts. In *The Twenty-Eighth International Flairs Conference*.

[31] Bhatnagar, S., Jacob, G., Devkar, G., Rybkowski, Z. K., Arefazar, Y., & Obulam, R. (2022). A systematic review of lean simulation games in the construction industry. *Architectural Engineering and Design Management*, 1-19.

[32] Gadre, S. Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., ... & Schmidt, L. (2023). DataComp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*.

[33] Hoffman, B. (2023) 3 Tactics for Outsmarting Artificial Intelligence, Psychology Today

[34] Altman, S. (2023), Wall Street Journal Tech Live

[35] Cropley, A. J. (2000). Defining and measuring creativity: Are creativity tests worth using?. *Roeper review*, *23*(2), 72-79.

[36] Kim, K. H. (2006). Can we trust creativity tests? A review of the Torrance Tests of Creative Thinking (TTCT). *Creativity research journal*, *18*(1), 3-14.

[37] Hu, W., & Adey, P. (2002). A scientific creativity test for secondary school students. *International Journal of Science Education*, *24*(4), 389-403.

[38] Brown, R. T. (1989). Creativity: what are we to measure?. In *Handbook of creativity* (pp. 3-32). Boston, MA: Springer US.

[39] Busov, B., Mann, D. L., & Jirman, P. (1999, May). TRIZ and invention machine: Methods and systems for creative engineering and education in the 21st century. In *proc. of the 1st International Conference on Advanced Engineering Design Methods, Prague*.

[40] Noever, D., Noever, S. preprint available online with Appendices,

AUTHORS

David Noever has research experience with NASA and the Department of Defense in machine learning and data mining. He received his BS from Princeton University and his Ph.D. from Oxford University as a Rhodes Scholar in theoretical physics.

Samantha Elizabeth Miller Noever has research experience in data science and social media analytics. She received her Bachelor's and Master's in Architecture from Catholic University of America, Washington, DC and taught at the Savannah College of Art and Design (SCAD).