
CEPHALO: MULTI-MODAL VISION-LANGUAGE MODELS FOR BIO-INSPIRED MATERIALS ANALYSIS AND DESIGN

Markus J. Buehler*

mbuehler@MIT.EDU

ABSTRACT

We present Cephalo, a series of multimodal vision large language models (V-LLMs) designed for materials science applications, integrating visual and linguistic data for enhanced understanding and interaction within human-AI and multi-agent AI frameworks. A key innovation of Cephalo is its advanced dataset generation method, which employs a sophisticated algorithm to accurately detect and separate images and their corresponding textual descriptions from Portable Document Format (PDF) documents, such as scientific papers. The method conducts a careful refinement of image-text pairs through integrated vision and language processing, ensuring high-quality, contextually relevant, and well reasoned training data. Cephalo is trained on integrated image and text data extracted from thousands of scientific papers and science-focused Wikipedia pages demonstrates can interpret complex visual scenes, generate precise language descriptions, and answer queries about images effectively. The combination of a vision encoder with an autoregressive transformer supports complex natural language understanding in an integrated model, which can be coupled with other generative methods to create an image-to-text-to-image or image-to-text-to-3D pipeline. To explore the development of larger models from smaller ones, we explore both the development of mixture-of-expert models and model merging. In model merging, we combine sets of layers that originate from different pre-trained source models. This hybrid approach allows us to leverage the domain-specific expertise and general conversational capabilities to harness the strengths of multiple models. We provide model weights for various sizes, ranging from 4 billion to 12 billion parameters, to accommodate different computational needs and applications. We examine the models in diverse use cases that incorporate biological materials, fracture and engineering analysis, protein biophysics, and bio-inspired design based on insect behavior. Other applications include the development of bio-inspired material microstructures, including pollen-based architected materials, as well as the synthesis of bio-inspired material microstructures from a photograph of a solar eclipse. We provide a detailed discussion of challenges and opportunities.

Keywords Language modeling · Multimodal generative AI · Scientific AI · Biomaterials · Bio-inspired materials · Inverse problems · Generative AI · Materials science · Multidisciplinary · Natural Language Processing · Computer vision · Materials informatics · Mixture-of-Experts Models

1 Introduction

In the rapidly evolving field of scientific artificial intelligence, and specifically as applied to materials science, approaches have explored a variety of scales, material types, and use cases [1, 2, 3, 4, 5, 6, 7, 8], integrating visual and linguistic data for advanced understanding and interaction has become an area of great interest. Applications include analysis of images, text or data mining. Emerging applications of such tools far exceed earlier computer vision methods [9, 10], such as image classification, as they provide more flexible and interactive methods of engaging with visual and text content. Moreover, the emergence of multi-agent AI systems [11, 12, 5] requires enhanced scientific vision capabilities to analyze data, assess generated plots, or interpret and reason over inputs from autonomous experimentation [13, 14].

*Massachusetts Institute of Technology (MIT), 77 Mass. Ave 1-165, Cambridge, MA 02139, USA

Such use cases of integrating different types of data, from text, to images, to content in scientific figures, tables and others, are critical especially in multidisciplinary areas of materials research such as bio-inspired materials, where researchers seek to integrate disparate sets of knowledge. Herein, generative AI is a particularly useful tool that can aid not only discovery but also finding engineering solutions for materials. This includes applications that include diverse properties that range from enhanced toughness to improved biological properties [15, 16, 17, 18, 19, 20, 21, 22, 8, 6].

A general framework of large language models (LLMs), here specifically to be understood as an approach that builds interaction graphs of complex input-output data with high capacity for in-context learning [23], has emerged as a promising approach in scientific analysis. Complex single and multi-modal models, such as those described in seminal works [24, 25, 26, 27, 28, 29, 5, 30, 31, 32], offer innovative pathways for knowledge expansion, especially when coupled with multimodal capabilities [33, 34, 5, 35]. The potential of LLMs in fostering new hypotheses is further supported by research in applications to specific domain tasks or general intelligence examinations [36, 37, 38, 39, 40, 41, 42, 34].

In this paper we focus on the development and application of Cephalo, a series of open-source multimodal vision large language models (V-LLMs), specifically in the context of bio-inspired design. The models are designed to bridge the gap between visual perception and language comprehension. Inspired by the intricate structures and mechanisms found in nature, particularly in bioinspired materials, Cephalo can interpret complex visual scenes and generate contextually accurate language descriptions. One of its key applications lies in analyzing and describing materials phenomena, such as failure and fracture, microstructures, and reasoning over biological and synthetic materials. Cephalo's architecture combines a vision encoder model with an autoregressive transformer to allow for tightly coupled visual and linguistic data processing. A variety of architectures and model sizes are explored to lay the foundation for a set of research-focused models.

1.1 Background and motivation

Integrating vision and language is a cornerstone in scientific research. In particular, materials science relies heavily on images and contextual analysis, ranging from understanding microstructures to interpreting data. In the field of bio-inspired materials, a key aspect is the translation of abstract concepts (structures to features to mechanisms) across fields, and visual cues are paramount. While language processing has advanced the capabilities of scientific AI in recent years, an important frontier is the establishment of models that can reason over a multitude of complex data that includes images and text, incorporating a set of complex scientific concepts. In this study, we explore the development and application of fine-tuned open source vision models, examined over a set of use cases that incorporate biological materials, fracture and engineering analysis, protein biophysics, and bio-inspired design. The study explores a variety of model architectures in different sizes and discusses how new models can be constructed by merging several smaller models into a larger one.

Transformer models learn during pre-training and fine-tuning. However, they also show superior ability of in-context learning, where models adapt their responses based on the context provided in the prompt. This context can include various forms of data, examples, or any relevant information. The ability of models to perform a wide range of tasks without task-specific training data or fine-tuning is a testament to the power of in-context learning. In earlier work LLMs have demonstrated the capacity to synthesize sophisticated understanding, such as translating between languages not included in their training [43]. Further, a recent article [44] discusses advancements in the interpretability of LLMs. The researchers used a sparse autoencoder to extract numerous interpretable features from the Claude 3 Sonnet model, a mid-sized LLM. These features range from abstract behavioral indicators to safety-relevant ones. The study highlights the potential of these features in enhancing our understanding of LLM behavior. Specifically, the mapping of features such as "Inner Conflict" indicates that the model learns and represents complex concepts. This is evidenced by several key factors. This includes, for instance, conceptual clusters, revealing the model's ability to organize features related to internal conflicts into distinct clusters such as balancing trade-offs, conflicting allegiances, and catch-22 scenarios suggests that it understands these concepts at a nuanced level. The authors [44] showed that each cluster represents a different aspect of internal conflict, showing the model's capability to differentiate and relate these aspects contextually. When features are activated or manipulated, the model's responses change accordingly. For instance, amplifying a feature related to the "Golden Gate Bridge" caused the model to focus excessively on that topic. Similarly, activating features related to inner conflicts altered how the model handled emotionally charged scenarios. This demonstrates that these features are not just random or superficial associations but deeply embedded representations that influence the model's understanding and output. The ability to represent, activate, and manipulate features associated with internal conflicts and other nuanced topics shows that the model has developed an internal understanding of these concepts, going beyond simple pattern matching to exhibit a form of conceptual learning and representation.

Other progress in the understanding of LLMs [42] proposes that representations in large LLMs are increasingly converging, where the authors first review numerous examples from literature showing that neural networks across

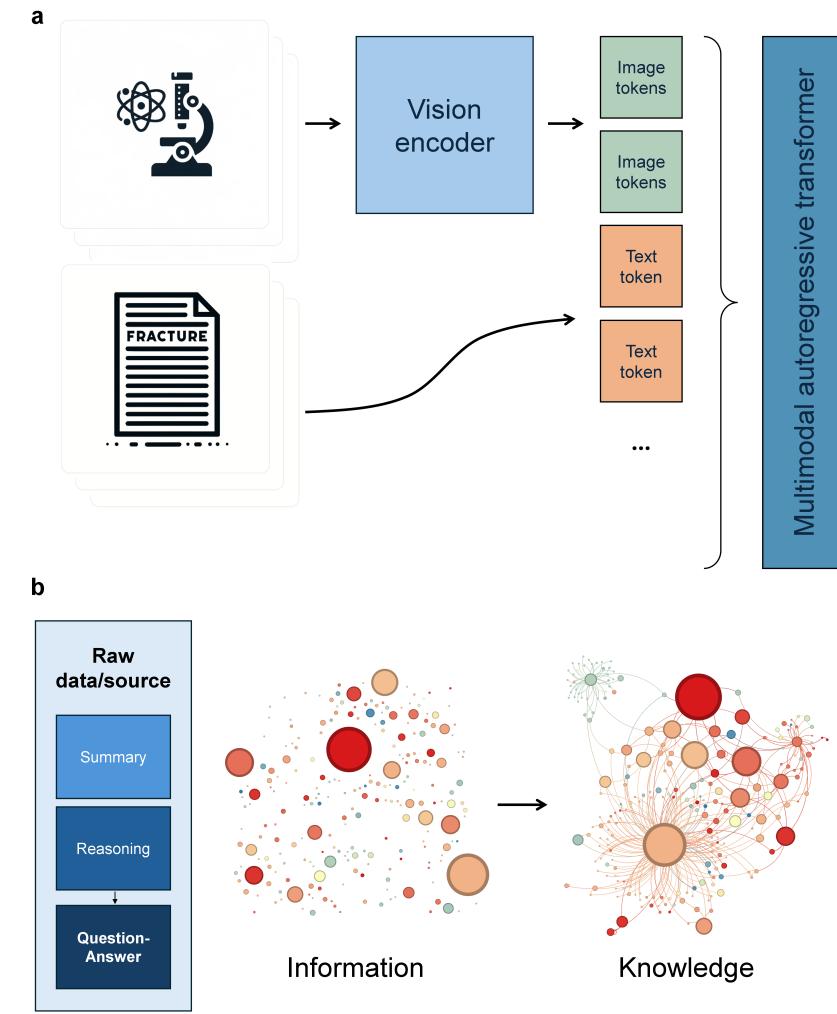


Figure 1: Overall approach used to develop the multi-modal vision LLM. Panel a: The model consists of a vision encoder that produces image tokens that are combined with text tokens in the autoregressive transformer model. Panel b: Delineation of the development of the dataset to train the model, consisting of both text-only data (taken from [39] and [37] as well as newly created image-text datasets).

different domains and over time are developing similar ways of representing data. They then demonstrate that as models grow larger, both vision and language models measure distances between data points in increasingly similar ways. The hypothesis suggests that this convergence is moving toward a shared statistical model of reality, analogous to Plato's concept of an ideal reality, termed the "platonic representation." The paper explores various selective pressures that might drive this convergence and discusses the implications, limitations, and counterexamples to their analysis.

These and other behaviors indicate that AI systems can be useful for knowledge discovery, especially when it comes to interrelating disparate areas of knowledge, drawing analogies, and predicting new insights from patterns in seemingly unrelated observations. We postulate that providing the proper context to facilitate discovery is essential for harnessing the full potential of these advanced AI systems, and that the incorporation of image data, combined with text and scientific principles, is a critical next step over previous approaches. While earlier work has focused on multimodal forward and inverse problems in the context of scientific applications [45, 34], here we generalize this strategy to develop reasoning abilities over diverse types of image-text combinations.

1.2 Outline of this paper

In this paper we present both the development of a series of vision models with a new dataset, as well as various applications to show the usefulness and applications in materials science, specifically focused on bio-inspired analysis and design. We provide an overview of the four models developed as part of this effort in Table 1.

The Cephalo model lineup <https://huggingface.co/1amm-mit/cephalo> consists of various versions designed for different purposes, three of which will be discussed in more detail in this paper. The base model, `Cephalo-Phi-3-vision-128k-4b-alpha`, trained on GPT-4o distilled data from Wikipedia and scientific papers, serves as a foundation. An improved counterpart, `Cephalo-Phi-3-vision-128k-4b-beta`, incorporates additional data from image-text data distilled using `Idefics-2` combined with a large text-only corpus, providing nuanced responses and excellent reasoning. These 4b class models are complemented with 8b and larger models. The `Cephalo-Idefics-2-vision-8b-alpha` model, trained solely on `Idefics-2` data, offers concise and generally accurate answers. Lastly, the `Cephalo-Idefics-2-vision-8b-beta` model, enhanced with GPT-4o distilled data, delivers longer, well-reasoned responses but can face challenges with complex concepts. We further use model merging can be used to develop larger V-LLMs at sizes 10b and 12b parameters by combining several smaller models. The `Cephalo-Idefics-2-vision-10b-alpha` model is a merged 10-billion-parameter model, featuring a total of 40 layers (32 original and 8 additional layers), with a checkpoint taken after the first epoch. It was trained on GPT-4o distilled image-text data sourced from Wikipedia and scientific papers. The `Cephalo-Idefics-2-vision-10b-beta` model, is based on a checkpoint taken after a second training epoch, trained on the same dataset. Additionally, the `1amm-mit/Cephalo-Idefics-2-vision-12b-alpha` model is a 12-billion-parameter merged model comprising 48 layers (32 original and 16 additional layers), with a checkpoint after the first epoch. This model was trained on a combined dataset derived from both `Idefics-2` and GPT-4o distillation of the paper corpus, including image-text data from Wikipedia and scientific papers.

We first provide a review of the basic architectures used here, then move to describe the dataset construction process, training, and development of inference strategies. The paper concludes with a series of case studies and a critical discussion, including an outline of future research in the field.

Model	Description	Training Data	Strengths/Weaknesses
<code>Cephalo-Phi-3-vision-128k-4b-alpha</code>	Base version of the Cephalo-Phi-3 model	GPT-4o distilled image-text data from Wikipedia and scientific papers	Good baseline model, but struggles in longer conversations. Context length of 128,000 tokens. Only one image per prompt.
<code>Cephalo-Phi-3-vision-128k-4b-beta</code>	Improved version of the Cephalo-Phi-3 model	GPT-4o and Idefics-2 distilled image-text data from Wikipedia and scientific papers, as well as a large text-only corpus	Provides nuanced responses, with excellent reasoning. Context length of 128,000 tokens. Only one image per prompt.
<code>Cephalo-Phi-3-MoE-vision-128k-3x4b-beta</code>	Mixture-of-expert model based on several smaller Cephalo-Phi-3 models.	GPT-4o and Idefics-2 distilled image-text data from Wikipedia and scientific papers, a large text-only corpus, as well image-to-LaTeX code dataset.	Diverse capabilities derived from the underlying expert models that make up the integrated model. Context length of 128,000 tokens. Only one image per prompt.
<code>Cephalo-Idefics-2-vision-8b-alpha</code>	Cephalo model based on Idefics-2	Idefics-2 distilled image-text data from Wikipedia and scientific papers	Gives shorter answers, to the point, and generally accurate. Handles multiple images per prompt.
<code>Cephalo-Idefics-2-vision-8b-beta</code>	Cephalo model based on Idefics-2	GPT-4o distilled image-text data from Wikipedia and scientific papers	Gives longer answers, with enhanced reasoning. Can struggle with complex concepts. Handles multiple images per prompt.
<code>Cephalo-Idefics-2-vision-10b-alpha</code>	Extended, larger/deeper Cephalo model based on Cephalo-Idefics-2-vision-8b-beta and Idefics-2	GPT-4o distilled image-text data from Wikipedia and scientific papers	Performs well overall. Handles multiple images per prompt.
<code>Cephalo-Idefics-2-vision-12b-alpha</code>	Further extended, larger/deeper Cephalo model based on Cephalo-Idefics-2-vision-8b-beta and Idefics-2	GPT-4o distilled image-text data from Wikipedia and scientific papers	Generally, does not perform as well as the 10b model. Handles multiple images per prompt.

Table 1: Summary of Cephalo models and their characteristics (models with sizes of 4b, 8b, 10b and 12b parameters are developed, including a mixture-of-expert model with 9b parameters total, and 4b active parameters). Please see <https://huggingface.co/1amm-mit/cephalo> for an overview of the models.

2 Results and discussion

The method can be applied to a variety of use cases, as summarized in Table 2.

We provide a brief discussion on dataset generation, albeit detailed methods are provided in the Materials and Methods section.

Use Case	Description	Implementation	Impact
Research Assistance	Literature Review and Data Extraction	Analyze scientific papers to extract key information and summarize findings on bioinspired materials.	Accelerates research by providing comprehensive literature reviews, saving time for researchers.
Image Analysis	Microscopic and Structural Analysis	Analyze images of natural materials to identify structural features for bioinspired material design.	Enhances understanding of natural structures, guiding the design of innovative materials.
Material Characterization	Comparative Analysis of Bioinspired Materials	Compare images and data of synthetic materials with their natural counterparts to assess design effectiveness.	Improves the evaluation process of new materials, ensuring they meet desired performance criteria.
Design Optimization	Suggesting Improvements in Material Design	Analyze current designs and suggest modifications to enhance material properties based on natural patterns.	Leads to the development of more efficient and effective bioinspired materials.
Educational Tool	Interactive Learning Platform	Create educational modules that use image-based learning to teach about bioinspired materials.	Facilitates learning and understanding of complex concepts in bioinspired material science.
Predictive Modeling	Predicting Material Properties	Predict properties of new materials by analyzing images and data from existing materials.	Accelerates the material development process by predicting properties before experimental validation.
Automated Documentation	Generating Reports and Visual Summaries	Generate detailed reports and visual summaries of research findings from experimental data and images.	Enhances communication of research findings, making it easier to share and publish results.
Innovation and Ideation	Brainstorming New Applications	Assist in brainstorming by providing insights and visual examples of bioinspired materials and their applications.	Sparks new ideas and applications, driving innovation in material science.
Quality Control	Monitoring Production Processes	Monitor production by analyzing real-time images to ensure quality and consistency of bioinspired materials.	Ensures high quality and consistency in the production of bioinspired materials, reducing defects.
Interdisciplinary Collaboration	Facilitating Cross-Disciplinary Research	Connect researchers from different fields by providing visual data and insights relevant to various disciplines.	Promotes innovative solutions and new discoveries through cross-disciplinary research and knowledge sharing.
New Material Discovery	Identifying Novel Materials	Use Cephalo to analyze vast datasets and images to discover new materials with unique properties.	Drives the discovery of novel materials, potentially leading to breakthroughs in various scientific fields.
Sustainability Analysis	Assessing Environmental Impact	Evaluate the environmental impact of bioinspired materials throughout their lifecycle.	Supports the development of sustainable materials and practices, contributing to environmental conservation.

Table 2: Summary of use cases for Cephalo in materials analysis, design, and related areas, along with their impact on materials science.

We first discuss dataset generation. Figure 2 depicts the overall approach to generate datasets for training the vision model. Reproductions of two representative pages of the scientific article (here, [46]).

Figure 3 shows histogram of the number of tokens for the image-text dataset, showing the source captions from Wikipedia (Figure 3) and the paper corpus (Figure 3b). Figure 3c-e show the results processed with different vision-text models. Figure 3c shows the histogram of the token numbers for the processed image descriptions for the Wikipedia (done using Idefics-2). Further, Figures 3d and e show the results for the paper corpus dataset, processed using Idefics-2 (Figure 3d) and GPT-4o (Figure 3e). The GPT-4o dataset generally yields much longer descriptions. A detailed analysis of the content shows that it provides much enhanced reasoning and nuanced explanation of the image content.

Figure 4 depicts a histogram of the image resolutions extracted from Wikipedia (Figure 4a) and the scientific paper corpus (Figure 4b), for X and Y directions, respectively, in the left and right columns of the plots.

One of the models is trained also on text-only data. For completeness we show an analysis of that dataset as well. Figure 5 shows a histogram of the number of tokens for the text-only dataset, showing questions only (Figure 5a), answers only (Figure 5b), and combined question-answer (Figure 5c). This dataset includes a corpus of knowledge extracted from scientific papers, books, and other sources in the area of biological materials, mechanics, and materials science.

To assess the quality of responses and capabilities for complex materials analysis tasks, we provide a series of experiments that are applied consistently for all models examined here. All examples include multi-turn interactions with a User (human), with consistent questioning for all cases studied. The specific cases include:

1. Analysis of a fracture scenario, where the model interacts with a human to assess two distinct fracture scenarios (brittle versus more ductile, distributed fracture), including an assessment of fracture mechanism and likely fracture properties of the material.

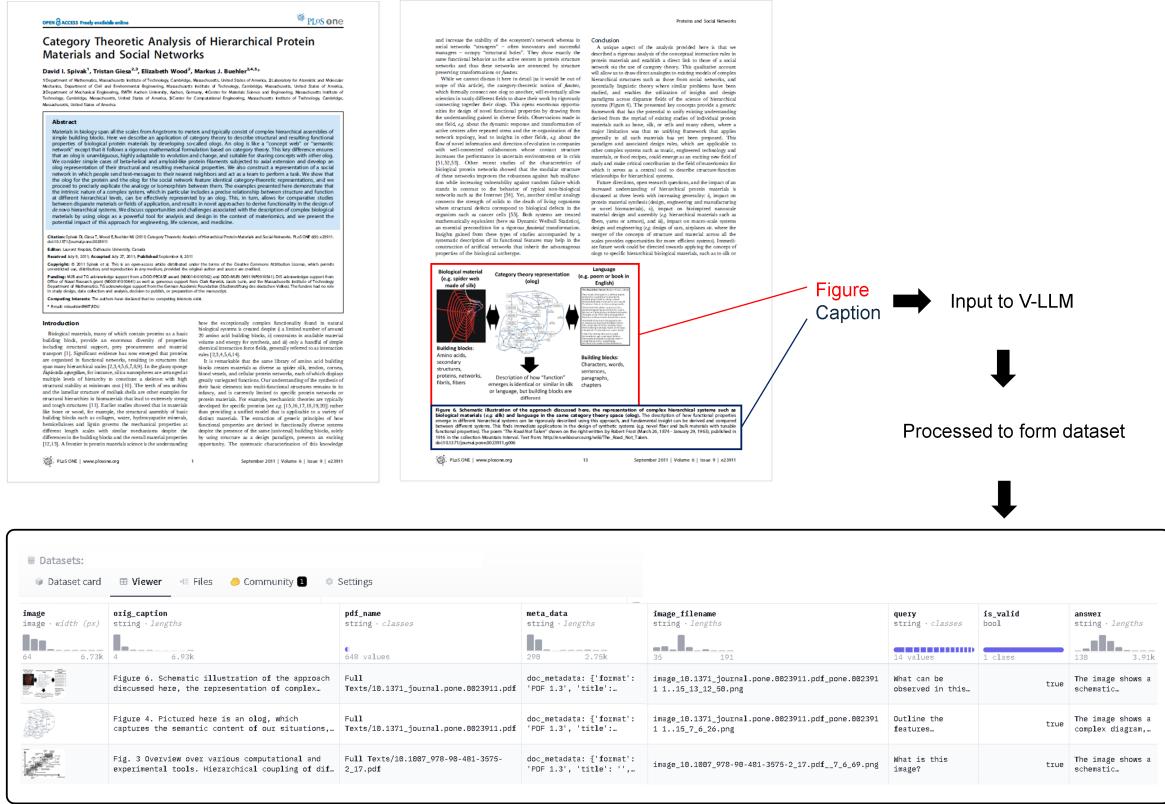


Figure 2: Visualization of the overall approach to generate datasets for training the vision model. Reproductions of two representative pages of the scientific article (here, [46], reproduced with permission from PLOS ONE via a Creative Commons License.

- Analysis of two proteins with distinct structures (highly organized beta-sheet protein on the one hand, and a partially disordered protein on the other hand). The model is tasked to assess the structure of the proteins based on their visualization, as well as estimate resulting mechanical properties.
- Analysis of an image of ants building an ant bridge via complex coordination, and assessing the content of the image in the context of applications in the development of multi-agent AI systems.
- Analysis of two images of distinct objects, including (1) an image of ants building an ant bridge via complex coordination, and (2) a micrograph of a pollen particle. The model is tasked to analyze both images and then develop bioinspired design concepts and research ideas based on them.

2.1 Cephalo-8b model series

We build our vision LLM based on a pretrained model, Idefics2. This base model is an 8-billion-parameter vision-language model that processes sequences of texts and images to generate text responses. It excels in tasks such as visual question answering, image description, document information extraction, and basic arithmetic. The baseline version offers OCR capabilities, the ability to manipulate images in their native resolutions and aspect ratios, and a simplified integration of visual features into the language backbone. The model utilizes a mixture of openly available datasets and is fine-tuned on task-oriented data, making it highly versatile for various multimodal applications. The parent models used are `google/siglip-so400m-patch14-384` and `mistralai/Mistral-7B-v0.1`.

The architecture processes images in their native resolutions and aspect ratios using the NaViT approach [48], avoiding resizing to fixed squares. Enhanced OCR capabilities and improved performance on charts, figures, and documents were achieved with targeted training data. The model does not use gated cross-attentions, and instead uses a vision encoder with Perceiver pooling [49] and MLP modality projection, concatenating visual features with text embeddings. The model was trained in multiple stages and leverages high-resolution images and various datasets, followed by instruction

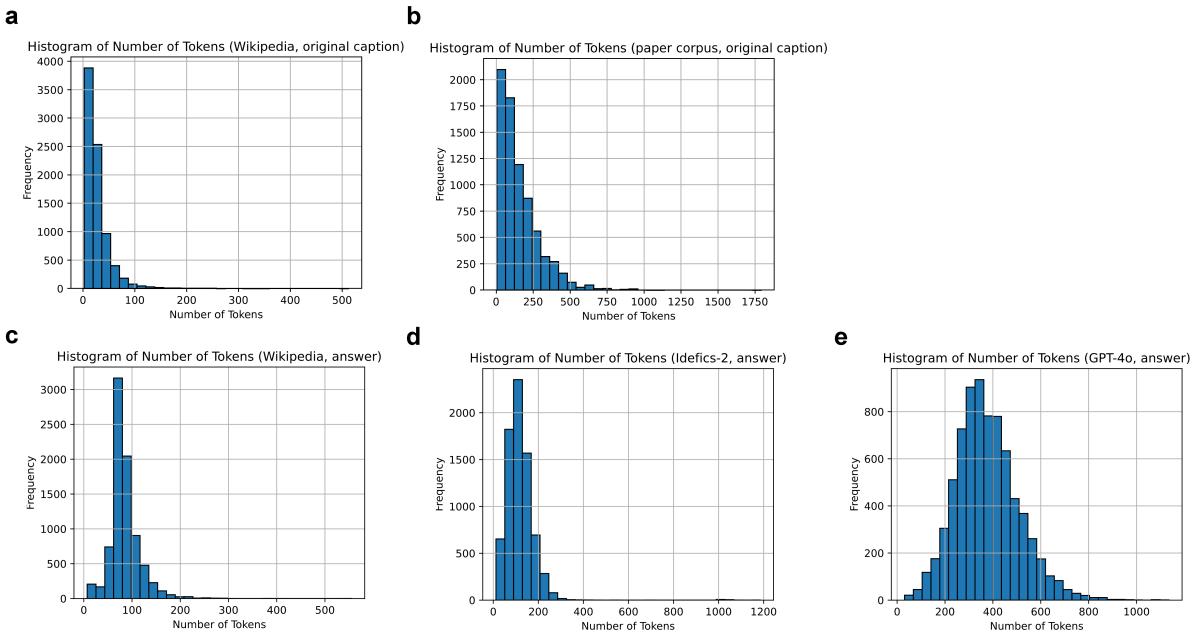


Figure 3: Histogram of the number of tokens for the image-text dataset, showing the source captions from Wikipedia (a) and the paper corpus (b). Panels c-e show the results processed with different vision-text models. Panel c shows the histogram of the token numbers for the processed image descriptions for the Wikipedia (done using Idefics-2). Panels d and e show the results for the paper corpus dataset, processed using Idefics-2 (d) and GPT-4o (e). The GPT-4o dataset generally yields much longer descriptions. A detailed analysis of the content shows that it provides much enhanced reasoning and nuanced explanation of the image content. All tokenization done using the Phi-3-Vision tokenizer [47].

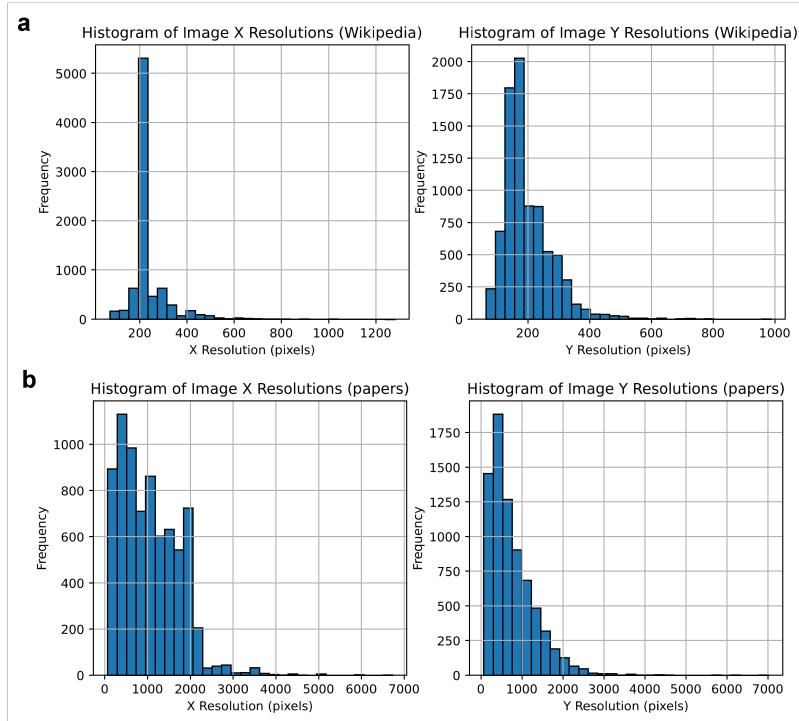


Figure 4: Histogram of the image resolutions extracted from Wikipedia (a) and the paper corpus (b), for X and Y directions, respectively (left/right column). All tokenization done using the Phi-3-Vision tokenizer [47].

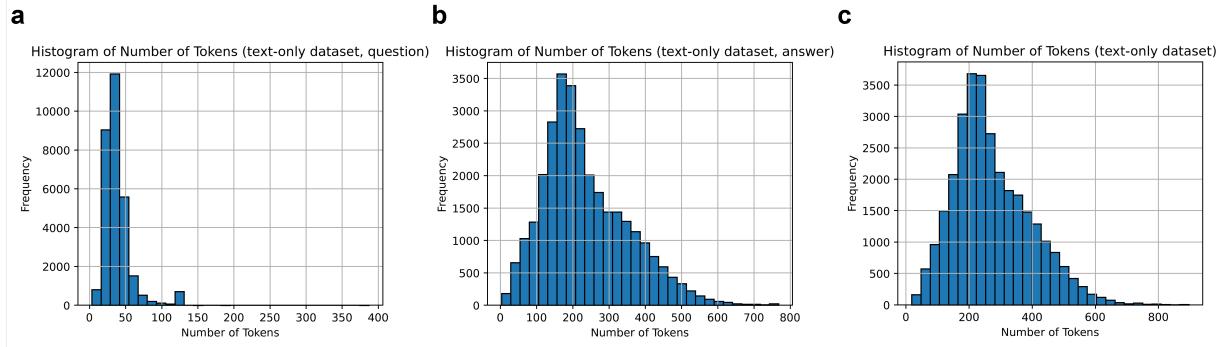


Figure 5: Histogram of the number of tokens for the text-only dataset, showing questions only (a), answers only (b), and combined question-answer (c). This dataset includes a corpus of knowledge extracted from scientific papers, books, and other sources in the area of biological materials, mechanics, and materials science. All tokenization done using the Phi-3-Vision tokenizer.

fine-tuning on curated vision-language datasets. In our work, we expand the fine-tuning to endow the model with enhanced capabilities to deal with scientific images.

As shown in 1a the model processes and integrates visual and textual data. It starts with the Vision Encoder, which processes input images to extract visual features. These features are then combined with text embeddings in the Vision-Language Connector using learned Perceiver pooling and MLP modality projection. This combined visual-text sequence is then fed into the LLM Decoder, which generates coherent text responses. This architecture thereby allows for effective handling of multimodal tasks such as visual question answering and document analysis.

The integration of Cephalo into the study and application of bioinspired materials presents numerous advantages, as summarized in Table 2. Key insights reveal that Cephalo can significantly accelerate research through automated literature reviews and data extraction, saving valuable time for researchers. Its ability to analyze microscopic and structural images enhances our understanding of natural materials, guiding the design of innovative bioinspired solutions. By predicting material properties and suggesting design optimizations, Cephalo streamlines the development process, leading to more efficient and effective materials. Additionally, its role in educational platforms facilitates learning, while its capabilities in generating reports and monitoring production processes ensure high-quality outcomes. Cephalo also fosters interdisciplinary collaboration by connecting researchers from different fields and providing insights relevant to various disciplines, driving innovative solutions and new scientific discoveries. Furthermore, its potential in identifying novel materials and assessing their sustainability impacts supports the development of breakthrough materials and environmentally conscious practices. Overall, Cephalo's application in bioinspired material science fosters innovation, improves evaluation processes, and enhances the quality and consistency of material production.

We look at two models, Cephalo-Idefics-2-vision-8b-alpha (based on Idefics-2 training data) and Cephalo-Idefics-2-vision-8b-beta (based on GPT-4o training data). Details about the training data and method of training is provided in the Materials and Methods section.

Both models perform well, but show different types of responses, and could be useful depending on required scenarios of application. Reflecting the nature of the training data with generally shorter responses (see, e.g. Figure 3d-e), Cephalo-Idefics-2-vision-8b-alpha provides relatively shorter replies. In contrast, Cephalo-Idefics-2-vision-8b-beta yields more elaborate responses. We find that sometimes, however, the model struggles to correctly capture logic and accuracy.

Text Boxes 1 (fracture mechanics), 2 (protein mechanics), 3 (bio-inspired AI systems) and 4 (bio-inspired materials and integration of disparate concepts) show results for the experiments conducted with Cephalo-Idefics-2-vision-8b-alpha.

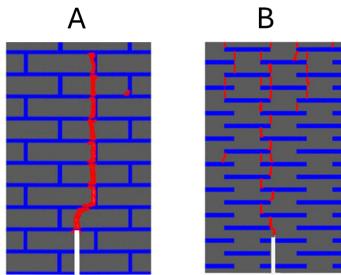
The conversation in Text Box 1 involves an analysis of two images (A and B) showing fracture scenarios. The images depict cracks in materials, with cracked areas in red and an initial white notch at the bottom. Cephalo is able to provide insights into the differences in crack propagation and suggests methods to toughen brittle composites. The discussion includes bio-inspired strategies and key insights are summarized in bullet points, as requested by the user. Key insights developed by the model include:

- The second image with multiple crack branches has a more distributed crack path than the first image with a single crack. This is correct, and represents a major difference between the two scenarios.

Chat between a User and Cephalo-Idefics-2-vision-8b-alpha

User: You are a materials scientist and you carefully analyze images, step by step.

Answer briefly.



User: You are shown two fracture scenarios.

Describe the differences in damage. Note, cracked material is shown in red color. The initial crack is the white notch at the bottom.

Cephalo: In the first image, the crack propagates in a straight line through the material, resulting in a single crack. In the second image, the crack splits into multiple branches, creating a complex network of cracks. The initial crack is the same white notch at the bottom in both images.

User: Which case has more distributed crack path?

Cephalo: The second image with multiple crack branches has a more distributed crack path compared to the first image with a single crack.

User: What steps could I use to toughen the brittle composite?

Cephalo: To toughen the brittle composite, several steps can be taken:

1. Introduce crack deflection mechanisms: This can be achieved by incorporating materials with different properties that can divert the crack path and reduce its size.
2. Enhance crack bridging: The use of fibers or particles can help bridge the crack and prevent it from growing.
3. Improve interfacial adhesion: Strong bonding between the phases can prevent crack propagation and improve toughness.
4. Increase crack-stopping mechanisms: Introducing materials that can absorb energy during deformation can help prevent crack growth and improve toughness.

By implementing these steps, the toughness of the brittle composite can be significantly improved.

User: Provide a bio-inspired strategy that uses honeycombs.

Cephalo: One bio-inspired strategy that uses honeycombs is the use of a hierarchical structure with multiple levels of organization. This approach mimics the natural structure of honeycombs, which provides high strength and toughness through a combination of crack deflection, crack bridging, and energy absorption mechanisms. The use of honeycombs can improve the toughness of brittle composites by providing a more distributed crack path and reducing crack growth.

User: Summarize key insights in a few bullet points.

- Cephalo:**
- The second image with multiple crack branches has a more distributed crack path compared to the first image with a single crack.
 - The second image is likely to have higher toughness due to the complex network of cracks that can dissipate more energy during deformation.
 - To improve the toughness of brittle composites, strategies such as introducing crack deflection mechanisms, enhancing crack bridging, improving interfacial adhesion, and increasing crack-stopping mechanisms can be implemented.

Text Box 1: Chat between a User and Cephalo-Idefics-2-vision-8b-alpha, focused on assessing two fracture scenarios. Image reproduced with permission from [50], using a Creative Common license (<http://creativecommons.org/licenses/by-nc-nd/3.0/>).

- The second image is likely to exhibit higher toughness due to the complex network of cracks that can dissipate more energy. The model correctly identifies this aspect.
- The model correctly identifies strategies to toughen brittle composites include introducing crack deflection mechanisms, enhancing crack bridging, improving interfacial adhesion, and increasing crack-stopping mechanisms.
- A bio-inspired strategy using honeycombs involves a hierarchical structure that mimics natural honeycombs for improved toughness.

It can be seen that Cephalo provides a clear comparison of the two fracture scenarios, highlighting the differences in crack propagation effectively. The model was able to make actionable suggestions. For instance, it suggests steps to toughen brittle composites that are practical and based on sound materials science principles. The honeycomb strategy is an innovative approach that ties biological concepts to materials. The key insights are summarized succinctly, making it easy to grasp the main points of the conversation. This shows a capability to accurately comprehend the conversation and successfully interact with the human user over multiple turns.

Chat between a User and Cephalo-Idefics-2-vision-8b-alpha

User: You are a materials scientist and you carefully analyze images, step by step.

Answer step-by-step.

A B

User: You are shown two proteins, colored by their secondary structure.

Looking at the protein images, and analyzing the structure, which of the proteins has a more organized structure?

Cephalo: Based on the image, it is evident that one protein has a more organized structure compared to the other. This can be attributed to various factors such as the presence of specific amino acid sequences, post-translational modifications, or environmental conditions that influence the protein's conformation and stability. The more organized structure may result in improved function or stability for the protein, depending on its specific role in the cellular context.

User: Based on this analysis, which protein (A or B) would have a higher mechanical resistance?

Cephalo: Based on the analysis, protein A is likely to have a higher mechanical resistance compared to protein B due to its more organized structure. This could be attributed to factors such as the presence of specific amino acid sequences, post-translational modifications, or environmental conditions that influence the protein's conformation and stability. The more organized structure may result in improved function or stability for the protein, depending on its specific role in the cellular context.

User: Suggest a few design improvements.

Cephalo: To improve the design of protein A, one could consider incorporating specific amino acid sequences or post-translational modifications that enhance its stability and mechanical resistance. Additionally, optimizing environmental conditions such as pH, temperature, or ionic strength during protein synthesis and assembly could also contribute to improved function and stability.

User: Be creative and think about using the proteins to generate a new electronic device. How would you go about it?

Cephalo: To create a new electronic device using proteins, one could consider incorporating these proteins into a conducting polymer matrix or a biocompatible substrate. This would allow for the formation of protein-based nanowires that can be used as electrodes or sensors in various applications such as biosensors, biofuel cells, or neural prosthetic devices. The use of proteins in electronic devices could potentially offer improved performance, biocompatibility, and functionality compared to traditional materials due to their unique properties and potential for self-assembly and molecular recognition.

Text Box 2: Chat between a User and Cephalo-Idefics-2-vision-8b-alpha, focused on assessing two proteins. The model is tasked to figure out structural features and to assess likely mechanical properties based on the images. The two proteins considered are PDB IDs 5B2A (crystal structure of a mutant of OspA, a beta-sheet rich structure with around 250 residues) and 1WVK (NMR solution structure of the partially disordered protein At2g23090 from *Arabidopsis thaliana*).

The conversation in Text Box 2 involves a user asking the model to analyze images of two proteins and evaluate their structural organization and mechanical resistance. The model provides insights into the more organized protein and suggests design improvements as well as creative applications for using proteins in electronic devices. The discussion focuses on the structural features and potential uses of the proteins. Key insights developed are as follows:

- Protein A has a more organized structure compared to protein B. This is an important aspect that was correctly identified by the model.
- The model correctly predicts that protein A is likely to have higher mechanical resistance due to its organized structure.

- When asked about design improvements, the model focuses on protein A. Specific ideas developed include incorporating specific amino acid sequences, post-translational modifications, and optimizing environmental conditions.
- A creative application involves using proteins in electronic devices by incorporating them into a conducting polymer matrix or biocompatible substrate to form protein-based nanowires.

Analyzing the conversation as a whole, we find that the model provides a clear comparison of the structural organization of the two proteins, effectively identifying the more organized protein. The design improvements for protein A are practical and based on sound principles of protein chemistry. The idea of using proteins in electronic devices shows creativity and an interdisciplinary approach, suggesting novel uses for biological materials in technology. As before, the key insights are summarized succinctly, making it easy to understand the main points of the conversation.

Chat between a User and Cephalo-Idefics-2-vision-8b-alpha

User: You are a materials scientist and you carefully analyze images, step by step.
Answer briefly.



User: Carefully study this photo.
What do you see?
Cephalo: The image shows a group of orange ants climbing up a black pole. The ants are moving in a line, with some ants leading the way and others following behind. The black pole is located in the center of the image, and the ants are scattered around it, some on the pole and others on the ground. The background is blurred, making the ants and the pole stand out.
User: How can the phenomenon be used to inform, or explain multi-agent AI systems?
Cephalo: The image of ants climbing up a pole can be used to inform or explain multi-agent AI systems by demonstrating how individual agents can work together to achieve a common goal. In this case, the ants are cooperating to climb the pole, with each ant contributing its own strength and expertise to the task. This behavior can be observed in multi-agent AI systems, where multiple agents work together to solve complex problems or complete tasks. The image provides a visual representation of the concept of collaboration and teamwork in multi-agent AI systems.
User: Provide a brief outlook to future research one could do.
Cephalo: One could conduct research on the biomechanics of ant climbing behavior to understand how they are able to climb vertical surfaces using their unique body structure and locomotion mechanisms. This could involve studying the ants' leg structure, adhesive properties, and sensory systems that enable them to climb efficiently. The findings from this research could inform the development of new robotics and artificial intelligence systems that mimic ant climbing behavior for various applications such as search and rescue operations or exploration of difficult terrains.
User: What type of deep learning model could be developed based on collective intelligence of ants?
Cephalo: A deep learning model could be developed based on the collective intelligence of ants by analyzing their behavior and interactions during the climbing process. This model would aim to mimic the ants' ability to cooperate and solve problems efficiently. The model could be trained using large datasets of ant climbing videos or images, which would provide valuable information about the ants' movements, interactions, and decision-making processes. By analyzing these data, researchers could identify patterns and trends in ant behavior that could be used to develop a deep learning model that mimics the ants' collective intelligence. This model could have potential applications in various fields such as robotics, artificial intelligence, and computer vision.

Text Box 3: Chat between a User and Cephalo-Idefics-2-vision-8b-alpha, analyzing an image of ants forming a bridge. Image reproduced from: <https://www.shutterstock.com/image-photo/unity-strength-1011406435>.

The next example shown in Text Box 3 focuses on analysis of an image of ants and relating it to multi-agent AI systems. The model describes the image and draws parallels between ant behavior and multi-agent AI systems, suggesting research directions and potential applications of a deep learning model based on collective intelligence of ants. The discussion highlights how the observed phenomenon can inform AI research and development. Key Insights include:

- The model correctly identifies the salient features in the image, in that it shows orange ants climbing a black pole, demonstrating teamwork and collaboration. However, it misses to understand the focus on an "ant bridge".
- The model correctly associates the phenomenon of insect collaboration as a means to be used to explain multi-agent AI systems where individual agents work together to achieve a common goal.

- The model proposes future research that could focus on the biomechanics of ant climbing behavior to inform robotics and AI system development.
- A deep learning model based on the collective intelligence of ants could be developed to mimic their cooperative problem-solving abilities.

The model provides an overall accurate description of the image, effectively highlighting the ants' collaborative behavior. The analogy between ant behavior and multi-agent AI systems is well-explained, making the concept accessible and relevant. The model's suggestions for future research and the development of a deep learning model are insightful, showcasing the potential for interdisciplinary applications in robotics and AI.

In the final experiment depicted in Text Box 4 the model is tasked to analyze two images: One showing ants near a tree trunk and a gap, and the other one showing a close-up view of pollen particle. Cephalo describes the images, compares their suitability for creating fracture-resistant materials, and suggests potential applications for the observed phenomena. The discussion progresses to developing a research idea that integrates the concepts from both images and provides a step-by-step outlook for future research. Summarizing the most important insights, we find:

- The model identifies correctly that the first image shows ants climbing a rough tree trunk, and associates with a demonstration of adaptation to challenging conditions, specifically material surfaces.
- The model identifies the second image as a plant cell wall with a repeating pattern of small holes, providing strength and stability. This is not entirely correct, as the image actually shows a pollen particle. Still, critical structural features are correctly identified.
- The model correctly identifies the second image on the right as being better suited for creating fracture-resistant materials due to its hard, rigid and porous structure.
- The model predicts that the ants' climbing behavior can be useful for studying locomotion and movement on rough surfaces.
- Future research is identified that could integrate the properties of plant cell walls and ants' adaptive movement to develop new materials with both strength and flexibility.

Overall, the model provides a clear and accurate description of both images, effectively highlighting the key features of the ants' climbing behavior and the plant cell wall structure. The analysis identifies the porous biological structure as better suited for creating fracture-resistant materials and suggests practical applications for the ants' behavior. The proposed research idea demonstrates creative thinking by integrating the properties of both images to inform the design of new materials. The step-by-step outlook for future research is well-structured and provides a clear road-map for further investigation.

We continue with the same experiment, but using the Cephalo-Idefics-2-vision-8b-beta model trained based on the GPT-4o based dataset. Text Boxes 5 (fracture mechanics), 6 (protein mechanics), 7 (bio-inspired AI systems) and 8 (bio-inspired materials and integration of disparate concepts) show results for the experiments conducted with Cephalo-Idefics-2-vision-8b-beta. Compared to the earlier examples, the responses of this model tend to be longer, more elaborate and feature better reasoning. This directly reflects the more sophisticated data used to train this model. The responses also feature better formatting (e.g., see Text Box 6 where bold typeset font is used in the enumerated list, using Markup language).

2.2 Model merging to create deeper, more expressive models: Cephalo-10b/12b model series

We construct a 10b Cephalo model by combining different parts of multiple models to create a new one [51, 52]. Model merging has been suggested as a powerful technique in machine learning and deep learning, albeit its use cases in vision-text models has not yet been explored. This approach leverages the strengths of each contributing model, often leading to improved performance and new capabilities that weren't present in the individual models. In our case, we use this strategy primarily to create a larger, deeper model, showcasing how individual components of pre-trained or fine-tuned models can be used to create new architectures.

The hybrid model Cephalo-Idefics-2-vision-10b-alpha demonstrated superior performance by effectively combining domain-specific expertise with general conversational capabilities. We merge the first 32 layers of the text_model decoder from the lamm-mit/Cephalo-Idefics-2-vision-8b-beta model with the last N_m layers of the chat/instruction-tuned HuggingFaceM4/idefics2-8b-chatty model. We fine-tune the last N_m layers of the merged model. Details are provided in the Materials and Methods section. For $N_m = 8$, this results in a 10b model. Figure 6 depicts a schematic of the approach.

Chat between a User and Cephalo-Idefics-2-vision-8b-alpha

User: You are a materials scientist and you carefully analyze images, step by step.

Answer concisely, and accurately.



User: You are shown two photos of two very different objects. Carefully study the two photos.

What do you see?

Cephalo: The first image shows a group of ants climbing up a tree trunk. The ants are moving in a line, with some ants leading the way and others following behind. The tree trunk is rough and textured, providing a challenging surface for the ants to climb.

The second image shows a close-up view of a cell wall in a plant. The cell wall is made up of small holes, which are arranged in a repeating pattern. The cell wall appears to be made of a material that is hard and rigid, providing support to the plant.

Both images show examples of materials interacting with their environment in different ways. The ants are adapting to the rough surface of the tree trunk, while the cell wall is designed to provide strength and stability to the plant.

User: Which of the two images would be better suited to create fracture-resistant materials?

Cephalo: The second image, showing the cell wall in a plant, would be better suited to create fracture-resistant materials. This is because the cell wall is made of a material that is hard and rigid, providing strength and stability to the plant. The repeating pattern of small holes in the cell wall also suggests a level of complexity and design that could be adapted for creating materials with improved fracture resistance.

User: What about the other one, what may it be useful for?

Cephalo: The first image, showing the ants climbing up a tree trunk, could be useful for studying animal locomotion and movement on rough surfaces. It may also provide insights into the ants' ability to adapt to challenging environments.

User: How can these concepts be used to inform new materials?

Cephalo: The concepts shown in these images can be used to inform new materials by studying the properties and structures of the materials in the images. For example, the cell wall in the plant could provide insights into the design of materials with improved fracture resistance, while the ants' ability to adapt to rough surfaces could inspire new materials that can move and interact with their environment in similar ways.

User: Next, specifically develop a research idea that integrates the concepts seen in the two images.

Cephalo: One potential research idea could be to study the cell walls of plants and their ability to provide strength and stability while also being flexible enough to adapt to challenging environments. This could involve analyzing the material properties of the cell walls and their structure, as well as testing their performance under different conditions. The ants' ability to climb up rough surfaces could also be incorporated into the research, with a focus on understanding how they adapt to their environment and how this could inform the design of new materials with improved fracture resistance and flexibility.

User: Provide a step-by-step outlook to future research.

Cephalo: The first step in future research would be to collect data on the cell walls of plants and their properties, as well as the ants' ability to climb up rough surfaces. This could involve conducting experiments or analyzing existing data sets. The next step would be to compare the properties of the cell walls with those of other materials, such as metals or polymers, to identify similarities and differences. Finally, the research could be used to inform the design of new materials that combine the strength and stability of the cell walls with the flexibility and adaptability of the ants' movement on rough surfaces.

User: Now write a detailed prompt for an image generation AI, that allows me to visualize the material. The prompt is:

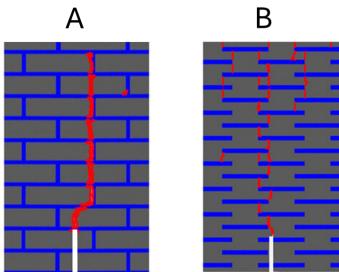
Cephalo: Visualize the material as a hard and rigid structure with a repeating pattern of small holes, similar to the cell wall in a plant. The material should also exhibit flexibility and adaptability, allowing it to move and interact with its environment in a similar way as the ants climbing up the rough surface.

Text Box 4: Chat between a User and Cephalo-Idefics-2-vision-8b-alpha, analyzing an image of ants forming a bridge and a microscopy image of a pollen particle. The model is tasked to reason over both and develop bio-inspired design concepts. Left image, reproduced from https://en.m.wikipedia.org/wiki/File:AntBridge_Crossing_10.jpg, based on a CC BY-SA 4.0, Igor Chuxlancev. Right image, reproduced from https://en.wikipedia.org/wiki/Pollen#/media/File:Arabis_voch1-4.jpg, CC BY-SA 3.0, Marie Majaura.

Chat between a User and Cephalo-Idefics-2-vision-8b-beta

User: You are a materials scientist and you carefully analyze images, step by step.

Answer briefly.



User: You are shown two fracture scenarios.

Describe the differences in damage. Note, cracked material is shown in red color. The initial crack is the white notch at the bottom.

Cephalo: The image shows two different fracture scenarios in a composite material.

In the first scenario (A), the crack propagates in a straight line, following the path of least resistance. The initial notch at the bottom is highlighted by a white line, and the crack extends vertically upwards, indicating a brittle fracture.

In the second scenario (B), the crack deviates from its initial straight path and follows a more complex, zigzag pattern. This suggests that the material has undergone significant plastic deformation before fracturing. The initial notch is again highlighted by a white line, but the crack path deviates from the straight line, indicating a more ductile fracture.

The results indicate that the material's behavior under stress is influenced by its microstructure and the presence of inclusions or other reinforcements. The straight crack in scenario A suggests a brittle fracture, where the material fails abruptly without significant plastic deformation. In contrast, the zigzag crack in scenario B indicates a ductile fracture, where the material undergoes significant plastic deformation before failing. This behavior is often observed in materials with a dual-phase microstructure, where the presence of a second phase (inclusions or reinforcements) can lead to crack deflection and energy dissipation, resulting in a more ductile fracture.

User: Estimate which scenario likely has higher toughness? Why?

Cephalo: Scenario B, with the zigzag crack pattern, likely has higher toughness compared to scenario A with the straight crack. This is because the zigzag crack path indicates that the material has undergone significant plastic deformation before fracturing. Plastic deformation absorbs energy, making the material more resistant to fracture. In contrast, the straight crack in scenario A suggests a brittle fracture, where the material fails abruptly without significant energy absorption, resulting in lower toughness.

User: Which case has more distributed crack path?

Cephalo: Scenario B has a more distributed crack path, as indicated by the zigzag pattern. This suggests that the material has undergone significant plastic deformation before fracturing, leading to a more complex crack path.

User: What steps could I use to toughen the brittle composite?

Cephalo: To toughen the brittle composite, various approaches can be employed, such as:

1. Introducing second-phase particles to create a dual-phase microstructure. These particles can deflect the crack path, leading to energy dissipation and improving the material's toughness.
2. Incorporating a gradient in the microstructure, where the material properties vary with position, can also enhance toughness by guiding the crack path and reducing the overall energy release rate.
3. Enhancing the interface between the phases to prevent premature crack initiation and propagation. This can be achieved through surface treatments or chemical modifications.
4. Controlling the processing conditions to achieve a more favorable microstructure, such as refining the grain size and optimizing the distribution of second-phase particles.

These methods aim to improve the material's ability to absorb energy and resist fracture, making the composite more resistant to cracking and tougher overall.

User: Provide a bio-inspired strategy that uses honeycombs.

Cephalo: A bio-inspired strategy using honeycombs involves mimicking the natural honeycomb structure found in some animal exoskeletons. This structure consists of a hexagonal lattice with varying cell sizes, providing both strength and flexibility. By replicating this design in composite materials, the honeycomb structure can enhance toughness by distributing stress more evenly and preventing premature crack initiation. This approach combines the benefits of both brittle and ductile fracture mechanisms, resulting in a material with improved overall toughness.

User: Summarize key insights in a few bullet points.

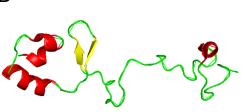
- The initial notch at the bottom of both scenarios indicates the starting point of the crack.
- Scenario A shows a straight crack path, suggesting a brittle fracture.
- Scenario B shows a zigzag crack path, indicating a more ductile fracture.
- The zigzag crack path in scenario B suggests that the material has undergone significant plastic deformation before fracturing.
- The straight crack in scenario A indicates a brittle fracture, where the material fails abruptly without significant energy absorption.
- The zigzag crack path in scenario B results in higher toughness compared to the straight crack in scenario A.

Text Box 5: Chat between a User and Cephalo-Idefics-2-vision-8b-beta, focused on assessing two fracture scenarios. Image reproduced with permission from [50], using a Creative Common license (<http://creativecommons.org/licenses/by-nc-nd/3.0/>).

Chat between a User and Cephalo-Idefics-2-vision-8b-beta

User: You are a materials scientist and you carefully analyze images, step by step.

Answer step-by-step.

A  B 

User: You are shown two proteins, colored by their secondary structure.

Looking at the protein images, and analyzing the structure, which of the proteins has a more organized structure?

Cephalo: The protein on the left has a more organized structure, as indicated by its detailed and complex arrangement of secondary structures. This suggests that the protein on the left is more stable and less prone to deformation under stress compared to the protein on the right, which appears to have a simpler and less organized structure.

User: Based on this analysis, which protein (A or B) would have a higher mechanical resistance?

Cephalo: Protein A would have a higher mechanical resistance due to its more complex and organized secondary structure, which provides greater stability and resistance to deformation under stress.

User: Suggest a few design improvements.

Cephalo: To enhance the mechanical properties of the protein, several design improvements can be made:

1. **Increase Hydrogen Bonding:** Introducing more hydrogen bonds within the protein structure can improve its stability and resistance to deformation.
2. **Increase Molecular Weight:** By increasing the molecular weight of the protein, its mechanical strength can be enhanced due to the increased number of atoms and bonds, providing greater resistance to deformation.
3. **Modify the Protein Sequence:** Altering the amino acid sequence of the protein can improve its mechanical properties by enhancing the stability of the secondary structures and reducing the likelihood of deformation under stress.
4. **Introduce Cross-Linking:** Adding cross-linking agents can help connect different protein chains, increasing the overall mechanical strength and resistance to deformation.

These design improvements would result in a protein with enhanced mechanical properties, making it more suitable for applications requiring high strength and stability.

User: Be creative and think about using the proteins to generate a new electronic device. How would you go about it?

Cephalo: To apply the proteins in creating an electronic device, we can consider using them as components in a molecular switch or sensor. The protein with a more organized structure (Protein A) could be used to fabricate a molecular switch, which can be controlled by external stimuli such as light or temperature changes. This switch could be integrated into a circuit to control the flow of electricity. Alternatively, the protein with a simpler structure (Protein B) could be used as a molecular sensor, detecting specific chemical or physical changes and transmitting signals to a larger device. This would involve modifying the protein to enhance its responsiveness to the desired stimuli.

Text Box 6: Chat between a User and Cephalo-Idefics-2-vision-8b-beta, focused on assessing two proteins. The model is tasked to figure out structural features and to assess likely mechanical properties based on the images. The two proteins considered are PDB IDs 5B2A (crystal structure of a mutant of OspA, a beta-sheet rich structure with around 250 residues) and 1WVK (NMR solution structure of the partially disordered protein At2g23090 from *Arabidopsis thaliana*).

In a transformer model, the earlier and later layers play distinct roles in processing input data. The earlier layers are primarily responsible for extracting low-level features and capturing local context. They utilize self-attention mechanisms to focus on immediate, neighboring tokens, identifying basic syntactic patterns, word identities, and short-range dependencies. As the data progresses through these layers, the model produces rich, contextually aware embeddings that encode fundamental aspects of the input.

Conversely, the later layers of a transformer synthesize higher-level abstractions and integrate global context. These layers leverage self-attention to consider tokens across the entire sequence, capturing long-range dependencies and complex relationships. By aggregating information from the entire input, the deeper layers can disambiguate meanings, resolve coreferences, and understand nuanced interactions. This hierarchical processing, moving from detailed token-level understanding to holistic comprehension in deep layers, enables the model to solve sophisticated tasks like contextual analysis of images, interactive human-AI conversations, text generation, summarization, translation and other tasks.

In terms of the modeling strategy used here, using a fine-tuned model for the early layers and a general-purpose model for the later layers leverages the strengths of both models effectively. The fine-tuned model, optimized for a specific domain, excels at extracting domain-specific features and understanding local context. These capabilities are critical in the initial layers, where precise and detailed feature extraction is necessary. In contrast, the general-purpose

Chat between a User and Cephalo-Idefics-2-vision-8b-beta

User: You are a materials scientist and you carefully analyze images, step by step.

Answer briefly.



User: Carefully study this photo.

What do you see?

Cephalo: The image shows a group of ants moving in a line along a surface. The ants are orange in color and appear to be following a specific path. They are positioned on a dark surface, which contrasts with their bright coloration. The ants are oriented in the same direction, indicating coordinated movement. The image captures the ants in a dynamic state, suggesting they are actively navigating or foraging. The background is blurred, focusing attention on the ants and their movement. This image likely illustrates the ants' ability to follow a specific path or trail, which is a common behavior observed in ant colonies.

User: How can the phenomenon be used to inform, or explain multi-agent AI systems?

Cephalo: The image of ants moving in a line can be used to inform or explain multi-agent AI systems by demonstrating how individual agents can coordinate their movements and follow a common path. This behavior is akin to how ants navigate through complex environments, following pheromone trails left by other ants. In multi-agent AI systems, this concept can be applied to develop algorithms that enable multiple agents to coordinate their actions and navigate through complex environments, mimicking the ants' ability to follow a shared path. The image highlights the importance of collective decision-making and coordination in multi-agent systems, which is crucial for achieving efficient and effective navigation in real-world applications.

User: Provide a brief outlook to future research one could do.

Cephalo: Future research could involve studying the mechanisms behind the ants' ability to follow a specific path. This could include investigating the chemical signals they use to communicate and coordinate their movements, as well as understanding how individual ants make decisions based on these signals. Additionally, research could explore the neural mechanisms underlying this behavior, potentially providing insights into how the brain processes information and makes decisions. This knowledge could be applied to develop more efficient and effective multi-agent AI systems that can navigate complex environments and coordinate their actions more effectively.

User: What type of deep learning model could be developed based on collective intelligence of ants?

Cephalo: A deep learning model based on the collective intelligence of ants could involve developing algorithms that mimic the ants' ability to navigate through complex environments by following pheromone trails. This model could incorporate principles such as reinforcement learning, where agents learn from past experiences and interactions, and distributed decision-making, where multiple agents work together to make collective decisions. The model would aim to replicate the ants' ability to coordinate their movements and navigate efficiently, which could be applied to various real-world applications such as autonomous robotics, search and rescue operations, and environmental monitoring.

Text Box 7: Chat between a User and Cephalo-Idefics-2-vision-8b-beta, analyzing an image of ants forming a bridge.
Image reproduced from: <https://www.shutterstock.com/image-photo/unity-strength-1011406435>.

model, trained on a broader dataset, brings strong generalization capabilities and a comprehensive understanding of diverse contexts, which are essential in the later layers. This combination ensures that the model retains the specialized knowledge of the fine-tuned model while benefiting from the general model's ability to integrate and synthesize information across varied multimodal contexts.

Deeper models tend to perform better generally because they can capture more complex patterns and relationships within the data. Each additional layer allows the model to process and refine the information further, moving from simple, local features to complex, global abstractions. This depth enables transformers to model intricate dependencies and interactions that are essential for high-level cognitive tasks. Moreover, the multi-head self-attention mechanism in each layer allows the model to attend to different parts of the input simultaneously, enhancing its capacity to learn diverse aspects of the data. The progressive refinement through deeper layers, from low-level detail extraction to high-level context integration, allows deep transformer models to achieve superior performance across a wide range of natural language processing tasks.

We show a few brief examples to showcase the performance of the model. Text Box 9 shows the results for the protein analysis problem. As can be seen from the responses, they are accurate and well defined, and follow clear logic without being too wordy. Another example is shown in Text Box 10, for the comparative analysis of ant dynamics and the microscale material. In a second experiment, we show the model images of an uncracked and a cracked egg, and task

Chat between a User and Cephalo-Idefics-2-vision-8b-beta

User: You are a materials scientist and you carefully analyze images, step by step.

Answer concisely, and accurately.



User: You are shown two photos of two very different objects. Carefully study the two photos.

What do you see?

Cephalo: The image shows two distinct objects:

1. The first object is a group of ants interacting with a tree branch. The ants are depicted in a natural setting, clustered together and moving along the surface of the tree branch. The ants appear to be working together, possibly foraging or building a nest.
2. The second object is a microscopic view of a biological structure, likely a cell or tissue sample. This image is magnified to reveal a detailed, porous structure with a honeycomb-like pattern. The scale bar indicates that the structure is on the microscale, with a size of 7 micrometers (μm).

These images illustrate the contrast between macroscopic and microscopic scales, highlighting the ants' behavior at the macro level and the detailed cellular structure at the micro level.

User: Which of the two images would be better suited to create fracture-resistant materials?

Cephalo: The second image, showing the detailed honeycomb-like structure, would be better suited for creating fracture-resistant materials. This is because the microscopic structure can provide insights into the mechanical properties and failure mechanisms of materials, which can inform the design of more durable and resistant materials. The macroscopic image of ants interacting with a tree branch does not provide direct information about material properties at the microscale.

User: What about the other one, what may it be useful for?

Cephalo: The first image of ants interacting with a tree branch could be useful for studying social behavior, ecology, or biomechanics. The ants' collective actions and interactions can provide insights into their communication, cooperation, and adaptability in their natural environment.

User: How can these concepts be used to inform new materials?

Cephalo: The concepts shown in these images can inspire the design of new materials by mimicking natural structures. The honeycomb-like structure observed in the second image can inform the development of materials with enhanced mechanical properties, such as improved fracture resistance and lightweight strength. The ants' collective behavior in the first image can inspire the design of materials that can adapt and respond to environmental stimuli, mimicking the ants' ability to work together and navigate their environment.

User: Next, specifically develop a research idea that integrates the concepts seen in the two images.

Cephalo: One research idea could involve studying the ants' collective behavior and the honeycomb-like structures in detail to gain insights into designing materials that can replicate these properties. By understanding how ants interact and communicate with each other and how their environment influences their behavior, researchers can develop materials that mimic these interactions, potentially enhancing their structural integrity and functionality. Similarly, by studying the honeycomb-like structures at the microscale, researchers can identify the key mechanical properties that contribute to their strength and durability, which can inform the design of new materials with similar characteristics.

User: Provide a step-by-step outlook to future research.

Cephalo: The next step in future research would involve further analysis of the ants' behavior and the honeycomb-like structures to understand their mechanical properties and adaptive responses. This could be achieved through detailed microscopic imaging, mechanical testing, and computational modeling to gain insights into the material's behavior under different conditions. The results from these studies could inform the development of new materials that mimic these natural structures, potentially leading to advancements in material science and engineering.

User: Now write a detailed prompt for an image generation AI, that allows me to visualize the material. The prompt is:

Cephalo: Show an image of a material with a honeycomb-like structure, highlighting its porous nature and the interconnectedness of its pores. The image should illustrate how this structure contributes to the material's mechanical properties, such as improved fracture resistance.

Text Box 8: Chat between a User and Cephalo-Idefics-2-vision-8b-beta, analyzing an image of ants forming a bridge and a microscopy image of a pollen particle. The model is tasked to reason over both and develop bio-inspired design concepts. Left image, reproduced from https://en.m.wikipedia.org/wiki/File:AntBridge_Crossing_10.jpg, based on a CC BY-SA 4.0, Igor Chuxlancev. Right image, reproduced from https://en.wikipedia.org/wiki/Pollen#/media/File:Arabis_vochl-4.jpg, CC BY-SA 3.0, Marie Majaura.

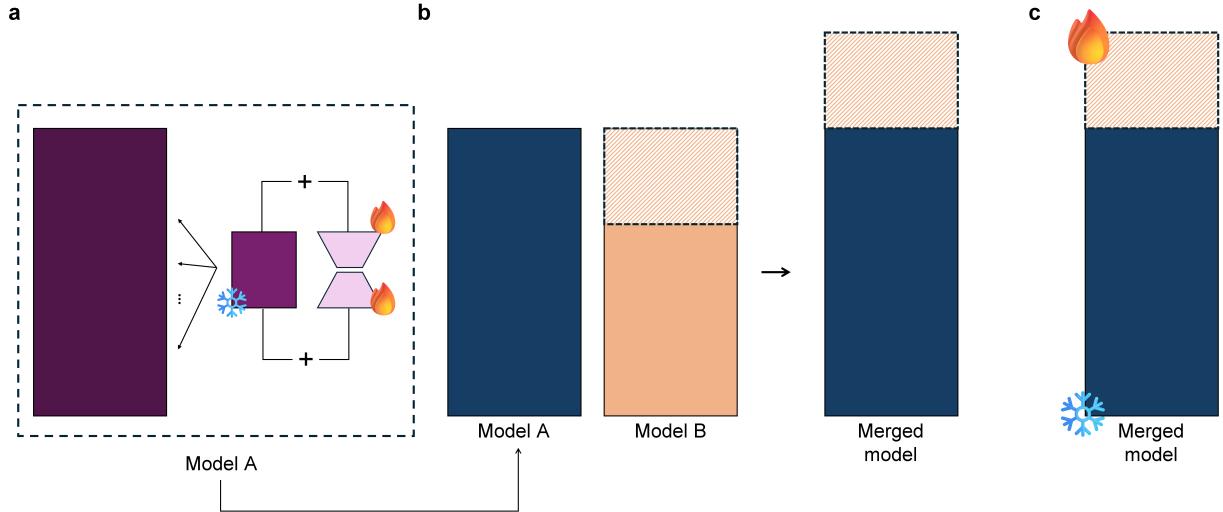


Figure 6: The development process of the merged Cephalo multimodal vision large language model (V-LLM). Panel a shows the process of fine-tuning the first model using low-rank adaptation [53]. Panel b: To merge, we use two models, Model A and Model B, to construct the larger model. Model A is a domain-specific fine-tuned model and Model B a general-purpose chat/instruction-tuned model. We select a set of layers (all layers from Model A, deep layers of Model B). This follows a strategy of using early layers from the domain-specific model and later layers from the general-purpose model. We then merge the selected layers into a new combined model, which is fine-tuned. Panel c: Fine-Tuning of the new model is done by freezing all layers that originate from Model A, and doing a full fine-tune of the layers that originated from Model B. The resulting model enables tasks such as image captioning, visual question answering, and multimodal content generation.

the model to assess the situation and predict what would happen if the object were to be dropped (see, Text Boxes 11 and 12).

This approach showcases the potential of merging models to harness the strengths of multiple models, leading to a versatile and powerful AI system, and to expand a model's depth by integrating more layers.

For a 12b merged model, we depict a few examples as well, in Supplementary Information. Generally, the 12b models do not perform as well as the 10b models, likely due to the fact that earlier layers of the second model are utilized more heavily in the process. For illustration, Text Box S1 shows the results for the protein analysis problem. Text Box S2 shows results for the comparative analysis of ant movement and the microscale material. Finally, Text Boxes S3 and S4 depict results of the experiments with images of an uncracked and a cracked egg.

2.3 Cephalo-4b model series

This version of Cephalo, [lamm-mit/Cephalo-Phi-3-vision-128k-4b-beta](https://huggingface.co/microsoft/Phi-3-vision-128k-instruct), is based on the Phi-3-Vision-128K-Instruct model [47]. The model has a context length of 128,000 tokens. Further details see <https://huggingface.co/microsoft/Phi-3-vision-128k-instruct> and [47]. Unlike the earlier models we use a more complex dataset in training, featuring text-only data (question-answer pairs from a corpus of biomaterials and mechanics of materials) combined with both the Idefics-2 based and GPT-4o based datasets.

Text Boxes 13 (fracture mechanics), 14 (protein mechanics), 15 (bio-inspired AI systems) and 16 (bio-inspired materials and integration of disparate concepts) summarize results for the experiments conducted with Cephalo-Phi-3-vision-128k-4b-beta. The responses of this model tend to be longer, more elaborate and feature better reasoning, and overall accurate. This directly reflects the more sophisticated data used to train this model, and shows the remarkable quality and performance of the Phi-series of models generally.

The performance of the model is very good overall, in spite of being of a smaller size than the Idefics-based 8b models. The presentation of the results is detailed, well reasoned, and accurate. For instance, Text Box 13 reveals a long but well articulated discussion of the fracture behavior. Text Box 13 reveals a long but well articulated discussion of the fracture behavior. Text Box 14 shows accurate and deep understanding of protein mechanics, connecting the visual representation of the structure with mechanical properties. The model focuses suggestions of improvement on protein B, which has less structural organization, and is thus the most likely candidate for such engineering tasks. The earlier two models did not identify this protein as an area of focus. As can be seen in Text Box 15 the model also provides a

Chat between a User and Cephalo-Idefics-2-vision-10b-alpha

User: You are a materials scientist and you carefully analyze images, step by step.

Answer step-by-step.

A B

User: You are shown two proteins, colored by their secondary structure.

Looking at the protein images, and analyzing the structure, which protein has a more organized structure?

Cephalo: The protein on the left has a more organized structure, as indicated by the presence of alpha-helices and beta-strands. These secondary structures contribute to the protein's stability and function. The protein on the right, in contrast, has a less organized structure with fewer defined secondary structures.

User: Which protein (A or B) would have a higher mechanical resistance?

Cephalo: Protein A would have a higher mechanical resistance due to its more organized structure, which includes alpha-helices and beta-strands. These secondary structures provide stability and strength to the protein, making it more resistant to mechanical forces.

User: Suggest a few design improvements.

Cephalo: To enhance the mechanical properties of the protein, several design improvements can be suggested:

1. Increase the number of alpha-helices and beta-strands in the protein structure. This would improve the overall stability and strength of the protein, making it more resistant to mechanical forces.
2. Introduce disulfide bonds between cysteine residues in the protein. These cross-links would further enhance the structural integrity and mechanical resistance of the protein.
3. Modify the amino acid sequence to optimize the formation of alpha-helices and beta-strands, ensuring a more organized and stable structure.
4. Introduce additional hydrogen bonds between amino acid residues, which would further stabilize the protein's structure and improve its mechanical properties.

By implementing these design improvements, the mechanical resistance of the protein can be significantly enhanced, making it more suitable for applications requiring robust materials.

User: Be creative and think about using the proteins to generate a new electronic device. How would you go about it?

Cephalo: To utilize the proteins for electronic device applications, the focus would be on leveraging their unique properties to create functional materials that can interact with electronic systems. This could involve incorporating these proteins into polymer matrices or composites to enhance specific properties such as conductivity, sensitivity, or mechanical strength. For example, the alpha-helical regions in the protein could be engineered to form conductive pathways, while the beta-strands could contribute to the mechanical stability of the material. This integration would allow for the creation of flexible, biocompatible electronic devices that can interact with biological systems, such as biosensors or implantable medical devices.

Text Box 9: Chat between a User and Cephalo-Idefics-2-vision-10b-alpha, focused on assessing two proteins. The model is tasked to figure out structural features and to assess likely mechanical properties based on the images. The two proteins considered are PDB IDs 5B2A (crystal structure of a mutant of OspA, a beta-sheet rich structure with around 250 residues) and 1WVK (NMR solution structure of the partially disordered protein At2g23090 from *Arabidopsis thaliana*).

detailed discussion of the ant image, correctly identifying bidirectional movement. It further goes into greater detail about the color of the insects. The remainder of the discussion is, like the others, lucid and well developed with good reasoning. In the final experiment shown in Text Box 16, the model correctly identifies the scenarios depicted in the two images - and, unlike the previous models, appropriately identifies the biological image as a "material with a porous structure". The step-by-step outlook to future research is detailed and accurate.

2.4 Image to text to image and 3D modalities

The previous section discussed human-AI interactions to reason over, and explore applications of, image-text modalities. Next, we show that by using Cephalo-Phi-3-vision-128k-4b-beta we can develop an integrated pipeline that features image-to-text-to-image modalities. The incorporation of generative image methods can thereby be used render a visualization of the designs developed by the model. The workflow involves two steps:

1. Reason over image-text associations, ending with (after one or more turns of interactions) the development of an image generation prompt.
2. Use of the image generation prompt developed by the Cephalo-Phi-3-vision-128k-4b-beta model to generate visualizations of novel images using a text-to-image or text-to-3D model.

Chat between a User and Cephalo-Idefics-2-vision-10b-alpha

User: You are a materials scientist and you carefully analyze images, step by step.

Answer concisely, and accurately.



User: You are shown two photos of two very different objects. Carefully study the two photos.

What do you see?

Cephalo: The image shows two distinct scientific observations:

1. **Ants Climbing a Surface**: The first image depicts ants climbing a surface. The ants, which are orange in color, are seen moving in a line, demonstrating their ability to climb vertical surfaces. This image highlights the ants' specialized structures and behaviors that enable them to defy gravity and move upwards.

2. **Nanoporous Material**: The second image shows a nanoporous material. This material is composed of a lattice structure with numerous small holes or pores. The image is magnified 15,000 times, revealing the intricate details of the material's porous architecture. The scale bar indicates that the image is at a microscopic level, with a width of 7 micrometers (μm).

The ants' climbing ability and the nanoporous material's structure are both examples of biological and material science, respectively. The ants' climbing behavior is a biological adaptation that allows them to move in environments where gravity would otherwise prevent them from doing so. The nanoporous material, on the other hand, is a synthetic material designed to mimic natural porous structures, which can have applications in fields such as catalysis, filtration, and energy storage.

User: Which of the two images would be better suited to create fracture-resistant materials?

Cephalo: The nanoporous material image would be better suited for creating fracture-resistant materials due to its intricate porous structure. The porous nature of the material can help in distributing stress and preventing cracks from propagating, making it a promising candidate for developing durable materials.

User: What about the other one, what may it be useful for?

Cephalo: The ants climbing image could be useful for studying biomimetic materials or developing new adhesive technologies. The ants' ability to climb vertical surfaces could inspire the design of materials that can adhere to surfaces in challenging environments, such as underwater or in space.

User: How can these concepts be used to inform new materials?

Cephalo: These concepts can inspire the design of new materials by combining biological principles with synthetic materials. For example, the nanoporous structure observed in the image can be replicated in synthetic materials to enhance their mechanical properties, such as strength and durability. Similarly, the ants' climbing behavior can inspire the development of adhesive materials that can stick to surfaces in various conditions. By combining biological insights with synthetic material design, researchers can create innovative solutions to practical problems in fields such as engineering, medicine, and environmental science.

User: Next, specifically develop a research idea that integrates the concepts seen in the two images.

Cephalo: The research idea could involve studying the mechanical properties of the nanoporous material and the ants' climbing behaviors to understand how these structures and biological adaptations contribute to their respective strengths. By examining the nanoporous material's stress distribution and the ants' adhesion mechanisms, researchers could aim to design materials that mimic these properties for enhanced durability and adherence in various applications.

User: Provide a step-by-step outlook to future research.

Cephalo: 1. **Nanoporous Materials**: Investigating the mechanical properties of nanoporous materials, such as their strength, flexibility, and resistance to fracture. This could involve studying how the porous structure influences stress distribution and potential applications in materials science.
2. **Ant Climbing Behavior**: Analyzing the ants' climbing mechanisms to understand how they adhere to surfaces and move against gravity. This could involve studying the anatomy of their feet and the biochemical properties that enable them to climb.
3. **Biomimetic Materials**: Developing new materials that mimic the ants' climbing abilities by incorporating similar adhesive properties into synthetic materials. This could involve creating materials with microscopic structures that allow them to adhere to surfaces in various conditions.

Overall, the research would focus on understanding the mechanical properties of nanoporous materials and the biological adaptations that enable ants to climb, with the goal of developing new materials that leverage these principles for enhanced durability and adherence in practical applications.

Text Box 10: Chat between a User and Cephalo-Idefics-2-vision-10b-alpha, analyzing an image of ants forming a bridge and a microscopy image of a pollen particle. The model is tasked to reason over both and develop bio-inspired design concepts. Left image, reproduced from https://en.m.wikipedia.org/wiki/File:AntBridge_Crossing_10.jpg, based on a CC BY-SA 4.0, Igor Chuxlancev. Right image, reproduced from https://en.wikipedia.org/wiki/Pollen#/media/File:Arabis_voch1-4.jpg, CC BY-SA 3.0, Marie Majaura.

Chat between a User and Cephalo-Idefics-2-vision-10b-alpha

User: You are a materials scientist and you carefully analyze images, step by step.

Answer concisely, and accurately.



User: Describe what you see. Do you detect a crack?

Cephalo: The image shows a single, whole, uncracked egg resting on a wooden surface. The egg's surface exhibits a natural, matte texture, typical of an unshelled egg. The background is a textured wood, providing a contrasting background to the smooth surface of the egg. There are no visible signs of cracking or damage on the egg's surface.

User: Use the image and propose what would happen if I were to drop the object.

Cephalo: If I were to drop the egg, it would likely crack due to the impact. The shell of an egg is composed primarily of calcium carbonate in the form of calcite, which provides structural integrity. However, under stress, such as from a drop, the shell can fracture, leading to the egg's contents being exposed. This is because the shell's structure is not continuous and is susceptible to damage when subjected to external forces.

Text Box 11: Chat between a User and Cephalo-Idefics-2-vision-10b-alpha, analyzing an image of an egg on a wooden surface. The model correctly states that there is no crack in the egg, and explains what would happen if the egg is dropped. The model provides a detailed description.

Chat between a User and Cephalo-Idefics-2-vision-10b-alpha

User: You are a materials scientist and you carefully analyze images, step by step.

Answer concisely, and accurately.



User: Describe what you see. Do you detect a crack?

Cephalo: The image shows a cracked eggshell on a wooden surface. The eggshell appears to have a rough texture and is partially broken, revealing the inner egg content. The cracks in the shell indicate stress points where the shell has given way, suggesting a point of impact or pressure. The wooden background provides a contrasting texture to highlight the eggshell's surface.

User: Use the image and propose what would happen if I were to drop the object.

Cephalo: If I were to drop the object, it would likely shatter into pieces due to the impact. The eggshell's brittleness would cause it to fracture along pre-existing cracks or new ones formed during the impact. This would result in the egg content being exposed and potentially damaged or spilled.

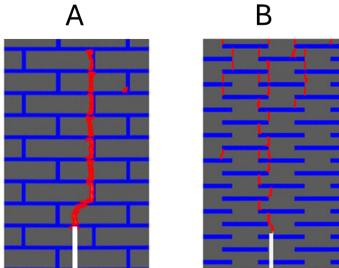
Text Box 12: Chat between a User and Cephalo-Idefics-2-vision-10b-alpha, analyzing an image of a pre-cracked egg on a wooden surface. The model correctly identifies the crack in the material and explains what would happen if the egg is dropped. Notable is a detailed description of what is seen in the image.

Chat between a User and Cephalo-Phi-3-vision-128k-4b-beta

User: You are a materials scientist and you carefully analyze images, step by step.

You are shown two fracture scenarios.

Describe the differences in damage. Note, cracked material is shown in red color. The initial crack is the white notch at the bottom.



Cephalo: In the image, two scenarios of material fracture are depicted.

In scenario A, the material fractures along a single line, forming a straight crack. This type of fracture is characterized by a clean, linear break in the material. The crack propagates from the bottom to the top of the image, indicating a unidirectional fracture path.

In scenario B, the material fractures along multiple lines, forming a branched crack pattern. This type of fracture is characterized by multiple cracks branching out from a central point. The cracks propagate from the bottom to the top of the image, indicating a multidirectional fracture path.

The differences in damage between the two scenarios are due to the different fracture mechanisms. In scenario A, the material fractures in a single direction, leading to a straight crack. In scenario B, the material fractures in multiple directions, leading to a branched crack pattern. The branched crack pattern in scenario B is more complex and can lead to a larger area of damage compared to the straight crack in scenario A.

User: Estimate which scenario likely has higher toughness? Why?

Cephalo: Scenario B likely has higher toughness. Toughness is a measure of a material's ability to absorb energy and plastically deform without fracturing. In scenario B, the material fractures along multiple lines, forming a branched crack pattern. This type of fracture is more complex and can absorb more energy before fracturing. In contrast, scenario A involves a single straight crack, which is simpler and can fracture more easily. Therefore, scenario B is likely to have higher toughness due to its ability to absorb more energy and deform plastically before fracturing.

User: Which case has more distributed crack path?

Cephalo: Scenario B has a more distributed crack path. In scenario B, the material fractures along multiple lines, forming a branched crack pattern. This type of fracture is characterized by multiple cracks branching out from a central point. In contrast, scenario A involves a single straight crack, which is simpler and more linear. Therefore, scenario B has a more distributed crack path due to the multiple cracks branching out from a central point.

User: What steps could I use to toughen the brittle composite?

Cephalo: To toughen the brittle composite, you could consider the following steps:

1. Introduce a toughening mechanism: One way to toughen the brittle composite is to introduce a toughening mechanism, such as crack deflection or crack bridging. This can help to dissipate energy and reduce the likelihood of fracture.
2. Modify the material properties: Another approach is to modify the material properties of the composite, such as its stiffness or ductility. For example, increasing the stiffness of the composite can help to reduce its tendency to crack under stress.
3. Improve the manufacturing process: The manufacturing process can also play a role in toughening the brittle composite. For example, using a more precise and controlled manufacturing process can help to reduce the likelihood of defects or imperfections that can lead to cracking.
4. Use a different material: If possible, consider using a different material that is more tough and less brittle. This could involve using a different type of composite material or a different type of material altogether.

It's important to note that the specific steps to toughen the brittle composite will depend on the specific application and the desired properties of the composite. Additionally, it's important to consider the trade-offs between toughness and other properties, such as strength and stiffness, when making modifications to the material.

User: Provide a bio-inspired strategy that uses honeycombs.

Cephalo: A bio-inspired strategy that uses honeycombs to toughen brittle composites is to mimic the structure and properties of honeycombs found in nature. Honeycombs are made up of hexagonal cells that are arranged in a specific pattern, which provides a high degree of toughness and resistance to cracking.

To apply this strategy to brittle composites, you could create a honeycomb-like structure within the composite material. This could involve using a specific type of material or manufacturing process to create the honeycomb structure. The honeycomb structure would help to dissipate energy and reduce the likelihood of cracking, similar to how honeycombs in nature provide toughness and resistance to cracking.

Another approach is to use a honeycomb-like structure as a toughening mechanism within the composite material. For example, you could introduce a toughening mechanism, such as crack deflection or crack bridging, within the honeycomb structure. This would help to dissipate energy and reduce the likelihood of fracture.

It's important to note that the specific approach to using honeycombs to toughen brittle composites will depend on the specific application and the desired properties of the composite. Additionally, it's important to consider the trade-offs between toughness and other properties, such as strength and stiffness, when making modifications to the material.

User: Summarize key insights in a few bullet points.

- Brittle composites can be toughened by introducing a toughening mechanism, such as crack deflection or crack bridging.
- The toughening mechanism can be achieved through a honeycomb-like structure within the composite material.
- The honeycomb structure can be created using a specific type of material or manufacturing process.
- The honeycomb structure can also be used as a toughening mechanism within the composite material.
- The specific approach to using honeycombs to toughen brittle composites will depend on the specific application and the desired properties of the composite.
- It's important to consider the trade-offs between toughness and other properties, such as strength and stiffness, when making modifications to the material.

Text Box 13: Chat between a User and Cephalo-Phi-3-vision-128k-4b-beta, focused on assessing two fracture scenarios. Image reproduced with permission from [50], using a Creative Common license (<http://creativecommons.org/licenses/by-nc-nd/3.0/>).

Chat between a User and Cephalo-Phi-3-vision-128k-4b-beta

User: You are a materials scientist and you carefully analyze images.

Answer step-by-step. You are shown two proteins, colored by their secondary structure.

Looking at the protein images, and analyzing the structure, which of the proteins has a more organized structure?

A



B



Cephalo: Protein A has a more organized structure than Protein B.

In the image, Protein A is shown to have a more regular and symmetrical arrangement of its secondary structure, with a consistent pattern of alpha-helices and beta-sheets. The yellow and green colors indicate the presence of these secondary structures, with the yellow representing alpha-helices and the green representing beta-sheets.

On the other hand, Protein B has a less organized structure, with a more random and irregular arrangement of its secondary structure. The red and green colors indicate the presence of these secondary structures, with the red representing alpha-helices and the green representing beta-sheets.

The more organized structure of Protein A suggests higher degree of stability and functionality, while the less organized structure of Protein B may indicate a lower degree of stability and functionality.

Therefore, based on the analysis of the secondary structure, Protein A has a more organized structure than Protein B.

User: Based on this analysis, which protein (A or B) would have a higher mechanical resistance?

Cephalo: Based on the analysis of the secondary structure, Protein A would have a higher mechanical resistance than Protein B.

The more organized structure of Protein A suggests a higher degree of stability and functionality, which would result in a higher mechanical resistance. The regular and symmetrical arrangement of the secondary structure in Protein A would provide a more consistent and predictable response to mechanical stress, making it more resistant to deformation and damage.

On the other hand, the less organized structure of Protein B would result in a lower degree of stability and functionality, and a more random and irregular response to mechanical stress. This would make Protein B less resistant to deformation and damage, and more likely to break or lose its functionality under mechanical stress.

Therefore, based on the analysis of the secondary structure, Protein A would have a higher mechanical resistance than Protein B.

User: Suggest a few design improvements.

Cephalo: Based on the analysis of the secondary structure, several design improvements could be suggested to enhance the mechanical resistance of Protein B.

1. Increase the regularity and symmetry of the secondary structure: By increasing the regularity and symmetry of the secondary structure, the mechanical resistance of Protein B could be enhanced. This could be achieved by optimizing the folding and assembly of the protein, or by introducing specific mutations that promote the formation of more stable secondary structures.

2. Enhance the stability of the secondary structure: The stability of the secondary structure is a key factor in determining the mechanical resistance of a protein. By enhancing the stability of the secondary structure, the mechanical resistance of Protein B could be improved. This could be achieved by introducing specific mutations that increase the stability of the secondary structure, or by optimizing the folding and assembly of the protein.

3. Optimize the interaction between secondary structures: The interaction between secondary structures is another important factor in determining the mechanical resistance of a protein. By optimizing the interaction between secondary structures, the mechanical resistance of Protein B could be improved. This could be achieved by introducing specific mutations that enhance the interaction between secondary structures, or by optimizing the folding and assembly of the protein.

4. Increase the size and strength of the protein: The size and strength of the protein are also important factors in determining its mechanical resistance. By increasing the size and strength of Protein B, its mechanical resistance could be improved. This could be achieved by introducing specific mutations that increase the size and strength of the protein, or by optimizing the folding and assembly of the protein.

Overall, these design improvements could enhance the mechanical resistance of Protein B, making it more resistant to deformation and damage under mechanical stress.

User: Be creative and think about using the proteins to generate a new electronic device. How would you go about it?

Cephalo: One potential way to use the proteins to generate a new electronic device is by exploiting their unique properties and structures. Here are some potential approaches:

1. Use the proteins as a substrate for electrochemical reactions: The proteins could be used as a substrate for electrochemical reactions, where their unique structures and properties could be leveraged to control the reactions and enhance their efficiency. For example, the proteins could be modified with specific functional groups or ligands that can facilitate the transfer of electrons between the reagents and the substrate.

2. Use the proteins as a catalyst for chemical reactions: The proteins could be used as a catalyst for chemical reactions, where their unique structures and properties could be leveraged to control the reactions and enhance their efficiency. For example, the proteins could be modified with specific catalytic sites or active sites that can facilitate the transfer of electrons between the reagents and the substrate.

3. Use the proteins as a semiconductor: The proteins could be used as a semiconductor, where their unique structures and properties could be exploited to control the flow of electricity between the electrodes. For example, the proteins could be modified with specific functional groups or ligands that can facilitate the transfer of electrons between the electrodes.

4. Use the proteins as a sensor: The proteins could be used as a sensor, where their unique structures and properties could be exploited to detect the presence of specific molecules or ions. For example, the proteins could be modified with specific binding sites or active sites that can selectively bind to specific molecules or ions, leading to changes in the electrical properties of the material.

Overall, these approaches could enable the use of the proteins as a new electronic device, with a wide range of potential applications in areas such as energy storage, sensing, and catalysis.

Text Box 14: Chat between a User and Cephalo-Phi-3-vision-128k-4b-beta, focused on assessing two fracture scenarios.

Chat between a User and Cephalo-Phi-3-vision-128k-4b-beta

User: You are a materials scientist and you carefully analyze images, step by step.

Answer briefly. Carefully study this photo.

What do you see?



Cephalo: The image shows a group of red ants, likely a species of fire ants, in a natural setting. The ants are clustered around a vertical wooden post, with some ants climbing up and others descending. The ants are primarily red in color, with some variations in shade and some individuals displaying a slightly darker hue. The post appears to be made of wood, with visible grain and texture. The background is a gradient of purple to pink, suggesting an outdoor environment. The ants are interacting with the post, possibly foraging or nesting.

User: How can the phenomenon be used to inform, or explain multi-agent AI systems?

Cephalo: The image shows a group of red ants, likely a species of fire ants, in a natural setting. The ants are clustered around a vertical wooden post, with some ants climbing up and others descending. The post appears to be made of wood, with visible grain and texture. The background is a gradient of purple to pink, suggesting an outdoor environment. The ants are interacting with the post, possibly foraging or nesting.

The image can be used to inform or explain multi-agent AI systems in several ways. First, it demonstrates the behavior of individual agents (ants) in a group setting, which is a key aspect of multi-agent systems. The ants are working together to achieve a common goal (foraging or nesting), which is similar to how multi-agent systems can be designed to work together to achieve a common objective.

Second, the image shows how individual agents can adapt and respond to their environment, which is an important aspect of AI systems. The ants are able to climb up and down the post, and adjust their behavior based on the position of the post and the presence of other ants. This adaptability is similar to how AI systems can be designed to respond to changes in their environment and adjust their behavior accordingly.

Third, the image shows how individual agents can interact with each other, which is another important aspect of multi-agent systems. The ants are working together to achieve a common goal, and their interactions with each other can affect the overall behavior of the group. This is similar to how AI systems can be designed to interact with each other and work together to achieve a common objective.

Overall, the image of red ants interacting with a vertical wooden post can be used to inform or explain multi-agent AI systems by demonstrating the behavior of individual agents in a group setting, their adaptability to their environment, and their interactions with each other.

User: Provide a brief outlook to future research one could do.

Cephalo: The image of red ants interacting with a vertical wooden post can be used to inform or explain future research in several ways.

One potential area of research could be the development of AI systems that are able to adapt and respond to changes in their environment, similar to how the ants in the image are able to adjust their behavior based on the position of the post and the presence of other ants. This could involve the development of algorithms that allow AI systems to learn from their environment and adjust their behavior accordingly.

Another potential area of research could be the development of AI systems that are able to interact with each other and work together to achieve a common objective, similar to how the ants in the image are working together to achieve a common goal. This could involve the development of algorithms that allow AI systems to communicate and coordinate their actions with each other.

Finally, the image of red ants interacting with a vertical wooden post could also inspire research into the design of AI systems that are able to work in natural environments, similar to how the ants in the image are able to work in a natural outdoor setting. This could involve the development of AI systems that are able to navigate and interact with their environment in a way that is similar to how the ants in the image are able to navigate and interact with their environment.

User: What type of deep learning model could be developed based on collective intelligence of ants?

Cephalo: The image of red ants interacting with a vertical wooden post could inspire the development of deep learning models that are based on the collective intelligence of ants.

One potential approach could be to develop a deep learning model that is able to learn from the behavior of individual ants and adapt its behavior based on the behavior of the group. This could involve the development of algorithms that allow the model to learn from the behavior of the ants in the image and adjust its behavior accordingly.

Another potential approach could be to develop a deep learning model that is able to interact with other models and work together to achieve a common objective. This could involve the development of algorithms that allow the model to communicate and coordinate its actions with other models, similar to how the ants in the image are working together to achieve a common goal.

Finally, the image of red ants interacting with a vertical wooden post could also inspire the development of deep learning models that are able to work in natural environments, similar to how the ants in the image are able to work in a natural outdoor setting. This could involve the development of AI systems that are able to navigate and interact with their environment in a way that is similar to how the ants in the image are able to navigate and interact with their environment.

Text Box 15: Chat between a User and Cephalo-Phi-3-vision-128k-4b-beta, focused on assessing two proteins. The model is tasked to figure out structural features and to assess likely mechanical properties based on the images. Image reproduced from: <https://www.shutterstock.com/image-photo/unity-strength-1011406435>.

Chat between a User and Cephalo-Phi-3-vision-128k-4b-beta

User: You are a materials scientist and you carefully analyze images, step by step.

Answer concisely, and accurately. You are shown two photos of two very different objects. Carefully study the two photos.

What do you see?



Cephalo: The image shows two distinct microscopic views of different materials. On the left, there is a close-up of a group of red ants on a tree branch. The ants are clustered together, with some individuals standing upright and others positioned horizontally. The background is blurred, emphasizing the ants and their immediate surroundings.

On the right, there is a microscopic view of a material with a unique structure. The material appears as a series of interconnected, porous spheres. The spheres are white and have a rough texture, with visible pores throughout. The image is in black and white, highlighting the contrast between the spheres and the background.

The left image is a photograph of ants, while the right image is a microscopic view of a material with a porous structure. The ants are red and are clustered together on a tree branch, while the material has a white, porous appearance.

User: Which of the two images would be better suited to create fracture-resistant materials?

Cephalo: The image on the right, showing a porous material, would be better suited to create fracture-resistant materials. The porous structure of the material can absorb and distribute stress, reducing the likelihood of fracture. The pores can also provide a pathway for cracks to propagate, further enhancing the material's fracture resistance.

User: What about the other one, what may it be useful for?

Cephalo: The image on the left, showing a group of red ants, may be useful for studying the behavior and social structure of ants. The ants are clustered together, indicating a social organization, and their red color may be a warning signal to potential predators. The image can provide insights into the behavior and communication of ants, as well as their role in the ecosystem.

User: How can these concepts be used to inform new materials?

Cephalo: The concepts shown in the images can be used to inform the development of new materials with specific properties. For example, the porous structure of the material on the right can inspire the creation of materials with enhanced fracture resistance. The social behavior of the ants on the left can inspire the development of materials with specific properties, such as self-healing or adaptive properties.

By studying the properties and behaviors of these materials, scientists can gain insights into how to design and create new materials with specific properties. For example, the porous structure of the material on the right can inspire the creation of materials with enhanced fracture resistance, while the social behavior of the ants on the left can inspire the development of materials with specific properties, such as self-healing or adaptive properties.

Overall, the concepts shown in the images can be used to inform the development of new materials with specific properties, and can provide insights into how to design and create materials with specific properties.

User: Next, specifically develop a research idea that integrates the concepts seen in the two images.

Cephalo: A possible research idea that integrates the concepts seen in the two images is the development of a fracture-resistant material inspired by the porous structure of the material on the right and the social behavior of the ants on the left.

The research idea would involve creating a material with a porous structure that can absorb and distribute stress, reducing the likelihood of fracture. The material would also have specific properties, such as self-healing or adaptive properties, inspired by the social behavior of the ants.

The research would involve studying the properties and behaviors of the ants and the porous material, and using this information to inform the development of the new material. The research would also involve testing the material's fracture resistance and other properties, and refining the material's design based on the results.

Overall, the research idea would involve integrating the concepts shown in the two images to create a new material with specific properties, and would provide insights into how to design and create materials with specific properties.

User: Provide a step-by-step outlook to future research.

Cephalo: A step-by-step outlook to future research on the development of a fracture-resistant material inspired by the porous structure of the material on the right and the social behavior of the ants on the left could include the following steps:

1. Conduct a literature review to identify existing research on porous materials and social behavior in ants.
2. Develop a research plan that outlines the objectives, methods, and timeline for the research.
3. Obtain funding and resources for the research, including equipment, materials, and personnel.
4. Conduct experiments to study the properties and behaviors of the ants and the porous material.
5. Analyze the results of the experiments and use this information to refine the design of the new material.
6. Test the fracture resistance and other properties of the new material.
7. Publish the results of the research in a scientific journal.
8. Use the results of the research to inform the development of new materials with specific properties.

Overall, the step-by-step outlook to future research would involve a systematic approach to developing a new material with specific properties, and would provide insights into how to design and create materials with specific properties.

Text Box 16: Chat between a User and Cephalo-Phi-3-vision-128k-4b-beta, analyzing an image of ants forming a bridge and a microscopy image of a pollen particle. The model is tasked to reason over both and develop bio-inspired design concepts. Left image, reproduced from https://en.m.wikipedia.org/wiki/File:AntBridge_Crossing_10.jpg, based on a CC BY-SA 4.0, Igor Chuxlancev. Right image, reproduced from https://en.wikipedia.org/wiki/Pollen#/media/File:Arabis_voch1-4.jpg, CC BY-SA 3.0, Marie Majaura.

In a first example we follow a similar human-AI interaction as in the previous section, but finish it off with the development of a prompt. The results are summarized in Text Box 17. The conversation involves a user asking the model to analyze two photos (same as before), and to develop design concepts for fracture resistant materials. The images show red ants near a branch and a microscopic porous biological structure. The model correctly identifies key fracture-resistant behaviors in both images. The ants exhibit a bridge-like structure that distributes stress, while the honeycomb pattern allows for efficient stress distribution and energy absorption.

The model proposes creating a new fracture-resistant bio-inspired microstructure by combining these principles. The suggested composite material features a layered design with an outer layer mimicking the ant bridge for stress distribution and an inner layer adopting the honeycomb pattern for energy absorption. It is suggested that this material would be synthesized using a layer-by-layer assembly technique, enhancing fracture resistance.

For the visualization task, the model provides a detailed prompt for an image generation AI to create a realistic depiction of the composite material. The prompt specifies a layered design with a bridge-like outer structure and a honeycomb inner pattern, using realistic colors to reflect the natural appearances of the ants and honeycomb. The resulting image should showcase the material's fracture-resistant properties in high resolution.

In the next experiment we automate this process and construct a pipeline that ingests an image, determines a prompt, and then renders the image using Stable Diffusion XL Turbo (SDXL-Turbo). We examine two examples, one where we start from the same microscopy image of a pollen particle as used before (Figure 7) and another one where we use an image of the total solar eclipse on April 7, 2024 (Figure 8). In both cases we task the model to reason over the image provided and bio-inspired design principles. As can be seen the vision model can successfully amalgamate, and then synthesize, information provided in the context as well as via the image provided. Notably, this works for both cases where the image is close to the design domain (that is, a micrograph of a pollen particle) or if the image is from a distinct domain (that is, an image of a solar eclipse).

The prompt used for the first example is:

Answer concisely, and accurately. Your task is to come up with an image generation prompt that combines concepts extracted from the cues in the image. The prompt must be brief, one sentence, but be a complete description of what the material looks like. I want to use the prompt for an image generation AI.

The prompt used for the second example is:

Answer concisely, and accurately. Your task is to come up with an image generation prompt that combines concepts extracted from the cues in the image with bio-inspired materials. The prompt must be brief, one sentence. Make sure the prompt describes a material microstructure but that it includes concepts from the image I am showing you. I want to use the prompt for an image generation AI.

A similar approach can be used also to develop 3D models, using generative text-to-3D strategies. Figure 9 depicts the results of one experiment, showing the design of 3D models from one of the prompts developed from the earlier examples, specifically combining the solar eclipse image with bio-inspired design. One can see that the design has similar features as the 2D images depicted in Figure 8, showing some level of consistency, as expected.

2.5 Mixture of Experts modeling: Constructing Large Models from Smaller Trained Component Models

Model merging, or generally building larger models is a powerful approach. However, the computational complexity and resource demands of these models often limit practical applications in materials science. To address these challenges, we propose an enhanced architecture that integrates a sparse mixture of experts (MoE) model [54, 55, 56], here implemented within the 4b model classes that are derived from the Phi-3-Vision architecture. Our approach leverages the strengths of multiple expert networks, dynamically allocating computational resources based on the input data. This not only improves the efficiency of the model but also enhances its performance on complex vision-language tasks. By adopting a sparse MoE mechanism, we can achieve diverse capabilities while maintaining practical feasibility. The Materials and Methods section (as well as the code implementation) present the details of the method, similar in approach as used in other recent AI models such as Mixtral [57]. Figure 10 shows an overview of the architecture.

This strategy has immense potential in future developments of multimodal vision models. For instance, we can construct a model with three experts derived from the original Phi-3-Vision-128k-4b model, Cephalo-Phi-3-vision-128k-4b-beta, and another fine-tuned model trained to convert images of equations into LaTeX code, Cephalo-LaTeX-Phi-3-vision-128k-4b-alpha (the model was trained on the image-to-LaTeX dataset, lammmit/Oleehy0-latex-formulas). This particular model with three experts has 4b active parameters at a given time (same as an original Phi-3-V model), but 9b parameters total. The sparse utilization of experts, here limited to $k = 1$

Chat between a User and Cephalo-Phi-3-vision-128k-4b-beta



User: You carefully analyze images, step by step.

Answer concisely, and accurately. You are shown two photos of two very different objects. Carefully study the two photos.

Use the concepts in both image and propose a new fracture-resistant bio-inspired microstructure for an advanced material that incorporates the information from both images.

Cephalo: The image shows two distinct microstructures: one is a group of red ants on a tree branch, and the other is a close-up of a microscopic structure resembling a honeycomb.

The red ants exhibit a natural fracture-resistant behavior. They form a bridge-like structure that distributes stress across multiple points, reducing the likelihood of catastrophic failure. This is akin to the concept of load-sharing in engineering, where stress is distributed to prevent localized failure.

The honeycomb structure, on the other hand, is a natural example of a fracture-resistant material. The hexagonal arrangement of cells allows for efficient stress distribution and energy absorption, similar to the principles of fracture mechanics.

To create a new fracture-resistant bio-inspired microstructure, we can combine the principles observed in both images. The bridge-like structure of the ants can be mimicked to distribute stress, while the honeycomb pattern can be replicated to enhance energy absorption.

A proposed microstructure could be a composite material with a layered design, where the outer layer mimics the ant bridge, and the inner layer adopts the honeycomb pattern. This would provide both stress distribution and energy absorption, making the material more resistant to fracture.

The material could be synthesized using a layer-by-layer assembly technique, where each layer is designed to mimic the properties of the respective structures. The result would be a material with enhanced fracture resistance, inspired by the natural designs observed in the images.

User: Now write a detailed prompt for an image generation AI like DALL-E or Stable Diffusion, that allows me to visualize the material.

The prompt must be brief, but be specific so that the image is rendered properly.

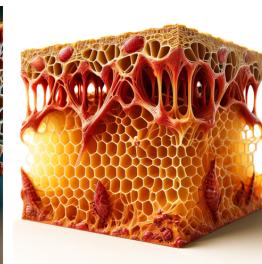
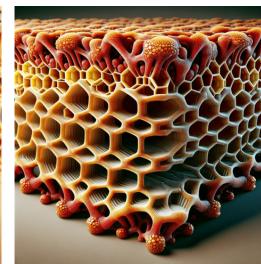
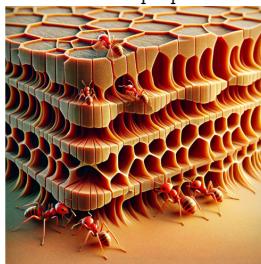
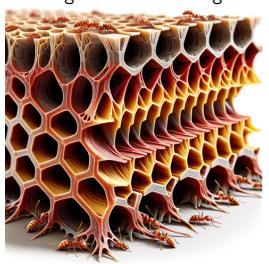
The prompt is:

Cephalo: "Create a detailed image of a composite material inspired by the fracture-resistant structures of red ants and honeycombs. The material should have a layered design with a bridge-like structure on the outer layer and a honeycomb pattern on the inner layer. Use a realistic color palette to represent the natural colors of the ants and the honeycomb structure. Ensure the image is clear and detailed, showcasing the fracture-resistant properties of the material."

The image generated by the AI should showcase the composite material with a clear and detailed representation of the bridge-like structure and honeycomb pattern. The colors should be realistic, reflecting the natural colors of the ants and the honeycomb structure. The material should appear to be fracture-resistant, with the bridge-like structure distributing stress and the honeycomb pattern absorbing energy.

The image should be rendered with high resolution to capture the intricate details of the material's structure. The AI should be able to interpret the prompt accurately and generate a realistic image that showcases the fracture-resistant properties of the material.

The resulting image should be a visual representation of the composite material's fracture-resistant properties, inspired by the natural designs of red ants and honeycombs. The image should be clear and detailed, allowing for a thorough understanding of the material's properties.



Text Box 17: Interaction between a User (human) and Cephalo-Phi-3-vision-128k-4b-beta to explore the development of an image generation prompt. The bottom of the text box shows four candidate images generated using GPT-4o. Left/top image, reproduced from https://en.m.wikipedia.org/wiki/File:AntBridge_Crossing_10.jpg, based on a CC BY-SA 4.0, Igor Chuxlancev. Right/top image, reproduced from https://en.wikipedia.org/wiki/Pollen#/media/File:Arabis_voch1-4.jpg, CC BY-SA 3.0, Marie Majaura.

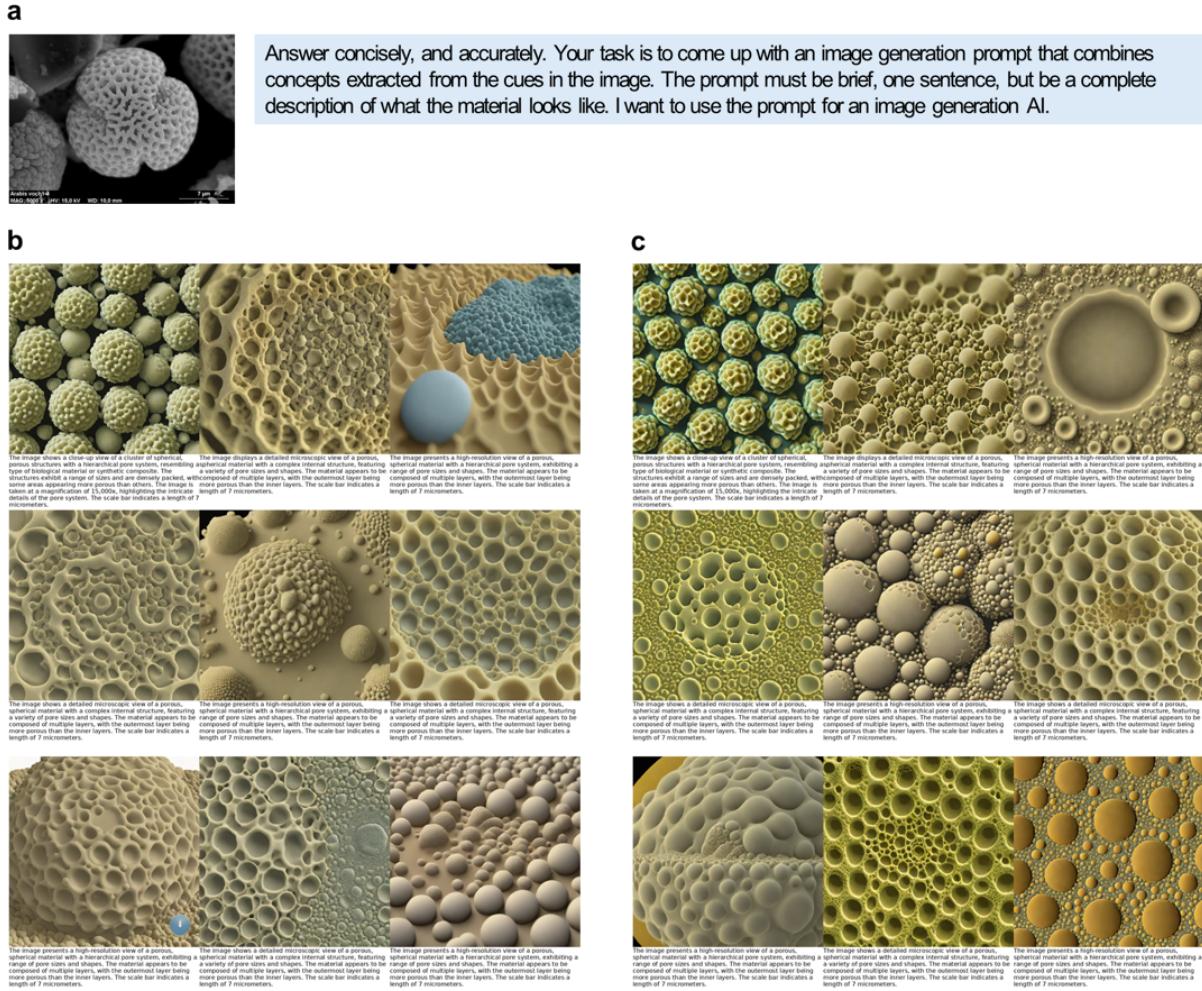


Figure 7: High-throughput image generation pathway. Panel a shows the original image, and panels b and c the results for 9 trials (panel b shows the results after two inference steps with SDXL-Turbo, and panel c the results after four inference steps). The respective image generation prompts so created are included as caption in each sub-panel of the figure. Left/top image, reproduced from https://en.wikipedia.org/wiki/Pollen#/media/File:Arabis_voch1-4.jpg, CC BY-SA 3.0, Marie Majaura.

expert, offers efficient performance at the cost of a single model but with access to much greater capabilities. In the scope of this paper, we present key details of our method, demonstrate its usefulness through a few initial experiments, and discuss its implications for future research. More work is left to future investigations. Text Boxes 18 and 19 show a few examples of how this model performs. The experiments show that the integrated model successfully integrates the capabilities of the constituting model. For reference, we render the LaTeX code as an equation here:

$$= 2(u+1)(2+t)b_1 + u(u+1)(b_1+c_1) + u(b_1+c_1+ub_3+uc_3) + 2u(u+1)c_2$$

The gating networks play a crucial role in determining the appropriate experts for each input token. These gating networks are trained using sample prompts that represent a diverse set of inputs. Such initial training can be complemented with further fine-tuning using complex datasets, and it is especially notable that this network can yield novel capabilities through the mixing of top k experts (each weighed with the weighting function obtained via a $\text{softmax}(\dots)$ activation function that produces a probability distribution over the top k experts.

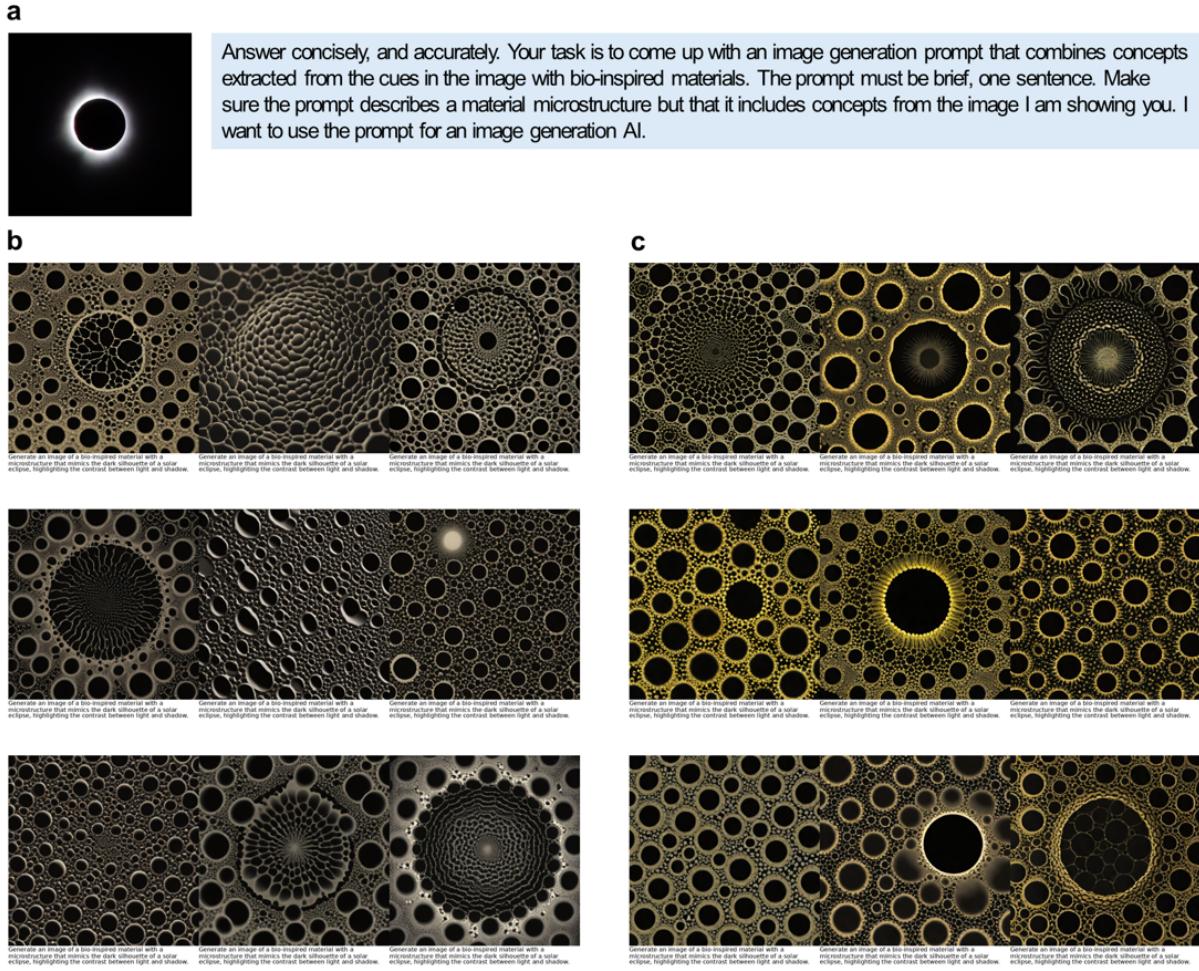


Figure 8: High-throughput image generation pathway. Panel a shows the original image, and panels b and c the results for 9 trials (panel b shows the results after two inference steps with SDXL-Turbo, and panel c the results after four inference steps). The image of the solar eclipse was taken in northern New England on April 7, 2024.

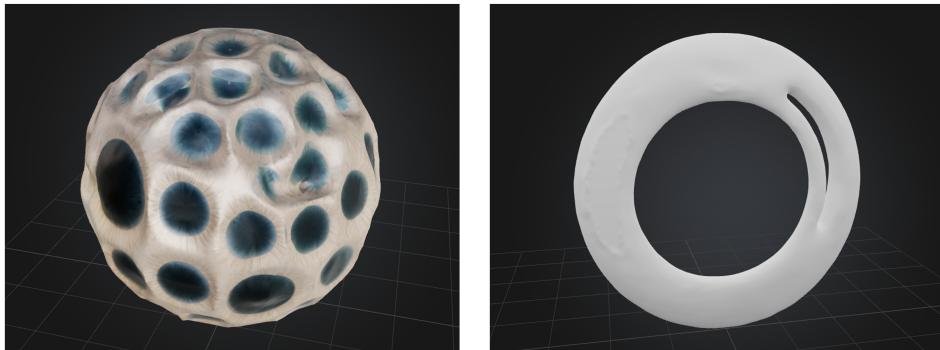


Figure 9: Use of a text-to-3D generative AI model to create a three-dimensional rendering of the prompt Generate an image of a bio-inspired material with a microstructure that mimics the dark silhouette of a solar eclipse, highlighting the contrast between light and shadow. using meshy.ai. Two variants are shown.

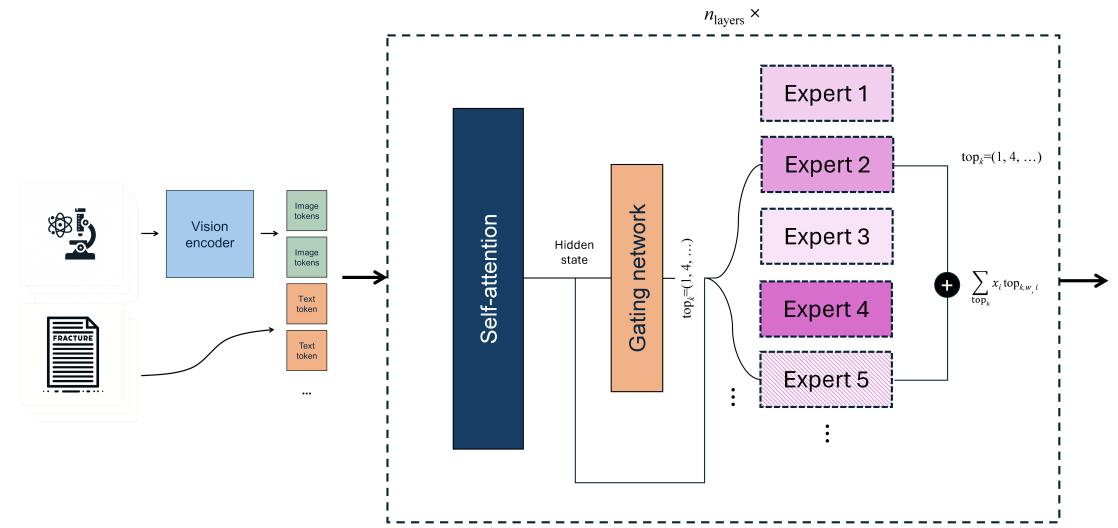


Figure 10: Development of a mixture-of-expert model with sparse gating. The components labeled “Expert i ” are extracted from a set of trained models, and hence inherit their capabilities. The input is similarly structured as in the original model, however, the inner mechanics of the model is altered. The schematic shows the overall workflow, where hidden states are routed through a gating network that decides which experts are activated. The results from top k experts are then added up, properly weighted. Normalization layers not shown for clarity.

Chat between a User and Cephalo-Phi-3-MoE-vision-128k-3x4b-beta

User: Provide a detailed answer.

What is shown in this image, and what is the relevance for materials design? Include a discussion of multi-agent AI.



Cephalo: The image depicts a group of ants, specifically fire ants, interacting with each other and their environment. The relevance for materials design lies in the study of how these ants build their nests and how their collective behavior can inspire the creation of new materials with unique properties. The ants' ability to work together and adapt to their surroundings can be applied to the development of materials that can self-assemble or change their properties in response to external stimuli.

In terms of multi-agent AI, the image highlights the importance of understanding and modeling the behavior of multiple agents working together towards a common goal. In the case of the ants, their collective behavior allows them to efficiently build their nests and protect their colony. By studying these interactions, researchers can develop algorithms and models that can be used to design and control multi-agent systems in various fields, such as robotics, logistics, and manufacturing.

In conclusion, the image of fire ants interacting with each other and their environment provides valuable insights into materials design and multi-agent AI. By understanding the collective behavior of these ants, researchers can develop new materials with unique properties and design algorithms and models for controlling multi-agent systems in various applications.

Text Box 18: Chat between a User and Cephalo-Phi-3-MoE-vision-128k-3x4b-beta, tasked to describe an image of ants building a bridge. Image reproduced from: <https://www.shutterstock.com/image-photo/unity-strength-1011406435>.

Chat between a User and Cephalo-Phi-3-MoE-vision-128k-3x4b-beta

User: Provide the LaTeX formula.

$$\begin{aligned} & 2(u+1)(u+2t+3)c_1 \\ & = 2(u+1)(2+t)b_1 + u(u+1)(b_1+c_1) + u(b_1+c_1+ub_3+uc_3) + 2u(u+1)c_2 \end{aligned}$$

Cephalo: \begin{aligned} & 2(u+1)(u+2t+3)c_1 \\ & = 2(u+1)(2+t)b_1 + u(u+1)(b_1+c_1) + u(b_1+c_1+ub_3+uc_3) + 2u(u+1)c_2 \end{aligned}

Text Box 19: Chat between a User and Cephalo-Phi-3-MoE-vision-128k-3x4b-beta, tasked to convert an image of an equation into LaTeX.

3 Conclusions

Cephalo is a series of multimodal vision LLMs (V-LLMs) that can reason over images, fine-tuned for biological and biologically-inspired materials. The model allows for a direct integration of visual and linguistic data for scientific research and practical applications, ranging from analysis to design to microstructure synthesis. Unlike traditional computer vision methods that focus, for instance, on specific classification tasks, the use of V-LLMs offers enhanced flexibility. The method can be used to perform a range of tasks with diverse difficulty, including classification tasks (e.g. does a crack exist) but can also integrate these skills with enhanced synthesis of information, new data, and step-by-step reasoning. This can either be used to develop complex insights or to develop new image generation prompts, closing the loop from image analysis to image synthesis, as shown in Text Box 17.

By leveraging an innovative dataset generation method that extracts high-quality image-text pairs from complex PDF documents as well Wikipedia scraping, Cephalo provides a framework for interpreting complex visual scenes and generating accurate, contextually relevant language descriptions. With future expansions of training data, additional evaluations and use cases, we anticipate that the models can be further improved. There are many immediate use cases, such as applications in autonomous experimentation, fine-tuning the models for specific applications (e.g. detection of cracks, damage, or identification of promising design features). The overall best performance is seen for the Cephalo-Phi-3-vision-128k-4b-beta model. The other models, albeit of larger size, struggle with some details in the images. However, these models have advantages, such as the ability to process high resolution images, and an exceptional ability to reason over multiple images.

From a fundamental perspective, its architecture, which combines a vision encoder with an autoregressive transformer, allows for sophisticated processing of diverse inputs, facilitating applications such as image captioning, visual question answering, and multimodal content generation. The deployment of Cephalo in the study of bioinspired materials, among other fields, underscores its potential to accelerate research, enhance understanding, and drive innovation. The capabilities not only push the boundaries of multimodal interaction and understanding but also offer a powerful tool for researchers to explore new frontiers in materials science and beyond. Among the four models developed, the Cephalo-Phi-3-vision-128k-4b-beta model exhibited the best performance, effectively integrating data from a more complex dataset to deliver nuanced and accurate outputs.

The models ranged in size from 4b to 12b and were constructed via several strategies. The use of model merging offers powerful avenues for future research, where models with different architecture, or distinct capabilities, can be developed. The concept of building more complex models from smaller ones was also be expanded via a mixture of expert modeling approach, here implemented for vision applications, such as in the Cephalo-Phi-3-MoE-vision-128k-3x4b-beta model. To do this we used a feed forward gating functions, but it could also be accomplished using low-rank based approaches (as done in X-LoRA [34]).

In several case studies, Cephalo's potential is illustrated:

- **Fracture Mechanics:** Cephalo accurately identified differences in fracture scenarios, providing insights into crack propagation and suggesting methods to enhance the toughness of brittle composites. Bio-inspired strategies, such as utilizing honeycomb structures, were proposed.
- **Protein Mechanics:** The model analyzed protein structures, predicting mechanical resistance based on visual cues and suggesting practical design improvements. Additionally, innovative applications, such as using proteins for electronic devices, were explored, showcasing Cephalo's interdisciplinary potential.
- **Multi-agent AI Systems:** By analyzing visual inputs of ants, Cephalo informed the development of multi-agent AI systems, emphasizing concepts of coordination and collective behavior. Future research directions, including the study of ant biomechanics to inspire robotics and AI systems, were also proposed.

- **Bio-inspired Materials:** The models integrated complex visual and textual data to develop new material designs, combining concepts from disparate sources like ant behavior and honeycomb structures. Detailed research plans and step-by-step approaches were generated, demonstrating the models' capability in aiding material science research.

A few other examples were explored, such as the detection of cracks in natural materials and structures, such as eggs. The mixture-of-expert model was used to demonstrate diverse capabilities that included the original bio-inspired focus, but also featured capabilities of the original Phi-3 model and a highly specialized version capable of translating images of equations to LaTeX code. This example illustrated the potential of using MoE strategies to accomplish highly complex, yet general, models.

The models were also used to facilitate a seamless pipeline from image-to-text-to-image generation, enhancing the visualization of novel material designs (see, e.g. Figures 7 and 8). Prompts generated by Cephalo were successfully used with models such as Stable Diffusion XL Turbo and Meshy (9) to create detailed 2D and 3D renderings of complex bio-inspired materials developed from a set of image cues. This capability underscores Cephalo's potential in visualizing and designing complex material structures, and the ability to reason over disparate concepts and cues over multiple interaction turns. This ability offers many use cases, especially in multi-agent AI applications. The development of a straightforward Mixture of Experts modeling strategy, with the code directly integrated into the model, provides high levels of flexibility. For one, it allows smaller fine-tuned models to be integrated. The ability to train the component models individually reduces cost and makes the development flexible, and easily useful, in scientific settings. The ability to integrate such small models into larger more functional models provides a scalable method to render more powerful models. While there are many use cases, the simple example reported here, of a model that has general knowledge, specialized bio-mechanics/bio-inspired materials knowledge, and a specialized capability to convert images of equations into LaTeX code can be useful to improve further dataset generation.

Further improvements can be made by expanding the training data and evaluating the models in additional use cases. Immediate applications include autonomous experimentation, further fine-tuning for specific tasks like damage detection, design improvements, or other uses cases, and the development of new bio-inspired designs that can be manufactured using 3D printing. The combination of visual and linguistic data processing renders the model as a powerful tool for interdisciplinary research, driving innovation in materials science and beyond. By leveraging the integrated vision-text transformer architecture and innovative dataset generation methods to extract detailed scientific information from a large corpus of literature, the model facilitates multimodal applications in research. One specific avenue that could be fruitful is to use one of the Cephalo models for further refinement of the training set, or to expand the dataset further. The improved understanding of scientific concepts will likely aid in developing more accurate, more nuanced and better reasoned training data. Other avenues can include the incorporation of video, which is especially important in understanding time-dependent phenomena.

4 Materials and methods

We summarize materials and methods used in this work.

4.1 Etymology and inspiration behind the name "Cephalo"

The name "Cephalo" is derived from the Greek word *kephalē* meaning "head" or "brain," which symbolizes the model's central role in processing and integrating visual and linguistic information. This name reflects the model's function as the "brain" of the system, facilitating advanced human-AI and multi-agent AI interactions through the comprehensive understanding of multimodal data. Additionally, "Cephalo" draws inspiration from cephalopods, a class of intelligent mollusks that includes octopuses, squids, and cuttlefish, associating it with the focus on biological inspiration that is central to the training and use of the model. Cephalopods are renowned for their exceptional cognitive abilities, advanced problem-solving skills, and highly developed nervous systems. They exhibit remarkable adaptability to their environments, sophisticated camouflage techniques, and complex behaviors, and are well-equipped to integrate visual cues with materialization. By naming our multimodal materials science V-LLM "Cephalo," we evoke the intelligence and adaptability of cephalopods. Similar to how cephalopods process diverse sensory inputs to navigate and respond to their surroundings, Cephalo integrates and processes visual and linguistic data to handle complex tasks. This dual inspiration highlights the model's potential for advanced problem-solving and contextual comprehension, drawing parallels between the cognitive prowess of cephalopods and the model's capabilities in the realm of materials science and beyond.

4.2 Dataset generation

Dataset generation is a key part, as scientific-focused vision models must be trained with focused and well-reasoned data that combines visual and text information than general-purpose datasets. To that end we develop a new dataset that incorporates data from Wikipedia, focused on a particular area of interest, with data extracted from scientific papers.

4.2.1 Data extraction from Wikipedia

We implemented a multimodal text/vision model for both inference and training, utilizing a dataset of images and LLM-processed data scraped from Wikipedia. The dataset generation code scrapes Wikipedia for images based on a set of search terms, leveraging generative AI to enhance the description of images for training purposes. This approach allows for the integration of complex logic in expanding image descriptions, thus enriching the training dataset. Once the image extraction algorithm is finalized, we will apply the same method to process figure-caption pairs, preparing the dataset for comprehensive testing and further model development.

We search for Wikipedia articles with these keywords:

```
Bioinspired materials, Mechanics, Mechanical engineering, Engineering, bone, spider, spider web, molecular modeling, dynamics, amino acid, hierarchical structure, seashell, Quantum mechanics, Molecular mechanics, Fluid dynamics, Solid mechanics, Mechanical properties of materials, Chemical engineering, Civil engineering structures, Electrical engineering innovations, Mechanical engineering design, Biomedical engineering, Biomimicry, Nanotechnology, Nanoscience, Biomaterials, Proteins, Biophysics, Biomaterials for medical applications, Biomimetic architecture, Bioinspired robotics, Sustainable materials from nature, Molecular biology, Cellular biology, Evolutionary biology, Biochemistry, Industrial chemistry
```

For each search keyword, we identify 100 highest hits using the Wikipedia API <https://en.wikipedia.org/w/api.php>. We access each of these sites using https://en.wikipedia.org/wiki/<SEARCH_HIT> and search each of these for image-caption pairs, by looping over all images. Then, we use the lamm-mit/Bioinspired-Phi-3-mini-4k to rewrite and expand on the original caption found in the Wikipedia site:

You follow directions. Do NOT repeat the question or task. Your job is to rewrite image descriptions in a precise way, using scientific principles. Your responses are concise but accurate, and include logic and reasoning.

Rewrite this description: `~~{caption}~~. Make sure that a complete description is provided, accurate and concise. Do NOT provide any figure or image number, citations, or references, just a clear description of what or who is shown.

Provide a succinct description, start with "The image shows..." or a variation thereof.

Figure 11 shows a screenshot of the Wikipedia image-text-dataset (see, lamm-mit/Cephalo-Wikipedia-Materials at Hugging Face). As can be seen, each image (left panel) is associated with a query-answer pair. The original caption is shown as well, and a comparison of it with the answer developed by lamm-mit/Bioinspired-Phi-3-mini-4k shows how the LLM adds important context to the description that goes beyond what was originally included in the Wikipedia pages. As one can confirm by exploring the dataset, captions on Wikipedia tend to be short and do not include full context, and hence, the additional processing into more elaborate question-answer pairs provides rich context and a higher quality training data. For completeness, and further additions or processing of the data, the dataset also includes references to the image URL, as well the URL of the Wikipedia article from which it was extracted.

The total size of the dataset includes around 7,500 image-text pairs.

4.2.2 Image-focused data extraction from a corpus of scientific papers

To develop a robust dataset generation method, we implemented a “from scratch” algorithm using the **fitz** library that is part of **PyMuPDF** <https://github.com/pymupdf/PyMuPDF>. The process begins by identifying all images on each page of a PDF. Subsequently, we locate text blocks that start with “Fig” or similar identifiers. The algorithm then matches these text blocks with the nearest image located below them. This matching process is refined through several clean-up steps, including handling different image colormaps and formats, and removing specific symbols, such as those added to documents by journals. The approach is robust, although some PDFs yielded divided figures, which may require additional processing to ensure dataset integrity.

Development of image-text pairs for training is done by sharing the image and original caption with a general-purpose V-LLM, and tasking the model to develop a comprehensive description of the image. We use both an open source V-LLM, **Idefics2**, and **GPT-4o**, for this task. As an alternative, we also explored using a text-only LLM (e.g., **Phi-3-Bioinspired**, which provides another option (akin to what was done for the Wikipedia dataset)). However, using vision-capable LLMs to process and distill datasets is generally better and provides more detailed and well-reasoned descriptions.

Datasets: lamm-mit Cephalo-Wikipedia-Materials					
Split(1) train					
The full dataset viewer is not available (click to read why). Only showing a preview of the rows.					
image	query	answer	orig_caption	image_url	url
	What is shown in...	The image displays the microstructure of sepiolite...	Sepiolite in solid form		https://en.wikipedia.org/wiki/Biomimetics
	What is shown in...	The image displays a cellulose nanocrystal...	Macroscopic picture of a film of cellulose...		https://en.wikipedia.org/wiki/Biomimetics
	What is shown in this image?	The image depicts the intricate microstructure of Morpho butterfly wings responsible for their vibrant blue coloration through structural coloration...	The vibrant blue color of Morpho butterfly wings responsible for their vibrant blue coloration through structural interaction with light. This natural optical property has inspired various technologies aiming to replicate it for potential applications in displays, sensors, and camouflage materials.		https://en.wikipedia.org/wiki/Biomimetics
	What is shown in...	The image displays rod-shaped tobacco mosaic virus...	Scanning electron micrograph of rod shaped tobacco mosaic...		https://en.wikipedia.org/wiki/Biomimetics
	What is shown in...	The image depicts the internal structure of a hip...	A hip implant is an example of an application of...		https://en.wikipedia.org/wiki/Biomaterial

Figure 11: Snapshot of the Wikipedia image-text-dataset (lamm-mit/Cephalo-Wikipedia-Materials at Hugging Face), revealing the organization of the data. Each image (left panel) is associated with a query-answer pair. The original caption is shown as well, and a comparison of it with the answer developed by lamm-mit/Bioinspired-Phi-3-mini-4k shows how the LLM adds important context to the description that goes beyond what was originally included in the Wikipedia pages (captions on Wikipedia tend to be short and do not include full context). The dataset also includes references to the image URL, as well the URL of the Wikipedia article from which it was extracted.

The process of constructing these follows this approach:

You follow directions. Do NOT repeat the question or task. Your job is to rewrite image descriptions in a precise way, using scientific principles. Your responses are concise but accurate, and include logic and reasoning.

Review this caption for the image: ``{caption}``. Now, rewrite it to state only facts, as a summary of what is shown in the image. Provide a detailed description of the image, starting with "The image shows...", "Shown in the image is..." or similar. Include details of content of what you can see in the image.

The response is:

Therein, {caption} denotes the caption extracted from the PDF file.

As a specific example, we show the image depicted in Fig. 12, including the original caption. The description of what is shown in the image obtained using Idefics2 is:

Idefics2

The image shows a longitudinal section through a pummelo (*C. maxima*) with the flavedo (a), albedo (b), and pulp (c) visible. A toluidine stained thin section of the peel is also present, showing the flavedo (a), albedo (b), remnants of the pulp (c), and vascular bundles (d). The tissue proportion [%] is plotted against the distance from the epidermis [mm].

For knowledge extraction from image-caption pairs using GPT-4o this we use a slight variation of the prompt, reflecting a far greater capability of the model to reason:

a

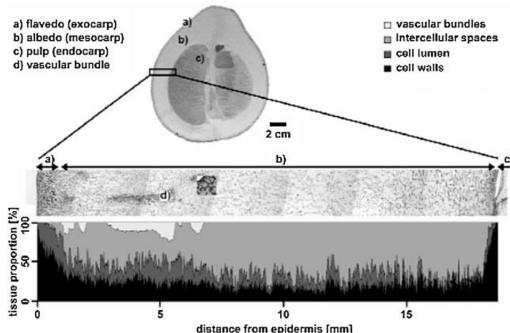


Fig. 2: Longitudinal section through a pummelo (*C. maxima*) showing flavedo (a), albedo (b), and pulp (c). Toluidine stained thin section of the peel showing flavedo (a), albedo (b), remnants of the pulp (c), and vascular bundles (d). Tissue proportion [%] plotted against the distance from the epidermis [mm].

b

The image shows a longitudinal section through a pummelo (*Citrus maxima*) depicting different layers of the fruit's peel and internal structures. The labeled and detailed sections include:

- **Flavedo (a) (exocarp):** This is the outermost layer, consisting of the colored part of the peel rich in oil glands.
- **Albedo (b) (mesocarp):** This is the white, spongy layer beneath the flavedo.
- **Pulp (c) (endocarp):** This refers to the edible, juicy segments inside the fruit.
- **Vascular Bundle (d):** These are part of the plant's vascular system, responsible for the transport of nutrients and water.

The magnified Toluidine-stained thin section of the peel shows more detail of these layers and structures. Toluidine blue is a dye used in microscopy for staining cell walls and other structures.

Additionally, a graph is included that plots tissue proportion (%) against the distance from the epidermis (measured in mm). The plot shows different levels ranging from the epidermis (0 mm) to 18 mm into the peel, indicating varying proportions of:

- **Vascular bundles:** Shown in light gray.
- **Intercellular spaces:** Shown in dark gray.
- **Cell lumen:** Areas inside the cells, shown in white.
- **Cell walls:** Shown in black.

The graph illustrates the relative proportions of these components as the distance from the outer peel increases. This provides insight into the structural composition and tissue distribution within the pummelo peel and inner sections. The varying proportions reflect the transition from the outer, denser exocarp, through the mesocarp, and into the vascular bundles integrated within these sections.

Figure 12: Example image and caption for extraction of image-text data. Panel a shows the original image and caption, and panel b the processed result that is used to form the training set. Image and caption reproduced with permission from [58]

You follow directions. Do NOT repeat the question or task. Your job is to rewrite image descriptions in a precise way, using scientific principles. Your responses are concise but accurate, and include logic and reasoning.

Carefully consider the image and the caption for the image: ``{caption}``.

Now, write a summary of what is shown in the image. State only facts, as a summary of what is shown in the image.

Provide a detailed description of the image, starting with "The image shows..", "Shown in the image is.." or similar.

Include details of the content of what you can see in the image. Define any terms, acronyms or specific technical words that are used.

If scientific results are shown, explain the logical reasoning behind the results.

Describe what the results shown in the image mean, if applicable.

The response is:

Using GPT-4o, we obtain:

GPT-4o

The image shows a longitudinal section through a pummelo (*Citrus maxima*) depicting different layers of the fruit's peel and internal structures. The labeled and detailed sections include:

- Flavedo (a) (exocarp): This is the outermost layer, consisting of the colored part of the peel rich in oil glands.
- Albedo (b) (mesocarp): This is the white, spongy layer beneath the flavedo.
- Pulp (c) (endocarp): This refers to the edible, juicy segments inside the fruit.
- Vascular Bundle (d): These are part of the plant's vascular system, responsible for the transport of nutrients and water.

The magnified Toluidine-stained thin section of the peel shows more detail of these layers and structures. Toluidine blue is a dye used in microscopy for staining cell walls and other structures.

Additionally, a graph is included that plots tissue proportion (%) against the distance from the epidermis (measured in mm). The plot shows different levels ranging from the epidermis (0 mm) to 18 mm into the peel, indicating varying proportions of:

- Vascular bundles: Shown in light gray.
- Intercellular spaces: Shown in dark gray.
- Cell lumen: Areas inside the cells, shown in white.
- Cell walls: Shown in black.

The graph illustrates the relative proportions of these components as the distance from the outer peel increases. This provides insight into the structural composition and tissue distribution within the pummelo peel and inner sections. The varying proportions reflect the transition from the outer, denser exocarp, through the mesocarp, and into the vascular bundles integrated within these sections.

Compared to the earlier response, this model is capable of providing a more well-developed description. Because the model is generally much more capable, it can provide nuanced data and accurately combine the caption with additional background knowledge and logical reasoning in the development of the description. The difference is also evident by looking at the token lengths, comparing answers derived using Idefics-2 and GPT-4o (Figure 3d-e).

The total size of each of the datasets developed from Idefics-2 and GPT-4o includes around 7,500 image-text pairs.

4.2.3 Text-focused dataset

To complement the image-text data, we also train some models with text-only data. The datasets are taken from [39] and [37] (lamm-mit/MechanicsMaterials), and cover a detailed training set from scientific papers in bioinspired materials, as well as mechanics and materials, especially failure mechanics. The dataset consists of around 30,000 question-answer pairs extracted from scientific papers and books.

4.3 Model training

The models are developed based on two pre-trained V-LLMs:

- Idefics2, an open-source multimodal model designed to process arbitrary sequences of image and text inputs to generate text outputs.
- Phi-3-Vision, an open-source multimodal model that can reason over image and text inputs to generate text outputs, with a large context length of 128,000 tokens.

The Idefics2 based models are trained on smaller datasets, focused on image-text data only. We found that the model struggled to learn well from combined text and image-text data. In contrast, the Phi-3-Vision based model trained well on the complex, integrated dataset and yielded a highly capable model overall.

In spite of these differences we find that all resulting models are versatile and can answer questions related to images, describe visual content, create narratives based on multiple images, or function purely as a language model without any visual inputs. They can also be used in complex pipelines that involve image analysis, reasoning, and generation of prompts for image generation in a multi-agent setup.

A summary of key training parameters is provided in Table 3 for both the 8b and 4b parameter models.

4.4 Prompt templates

The 8b models (as well as 10b and 12b models) use this prompt template:

Table 3: Training parameters used in the development of the models, including low-rank adapter (LoRA) parameters [53] (see also Figure 6a for a schematic.).

Parameter	8b Model	4b Model
Learning rate	1E-5	1E-5
Maximum gradient norm	0.5	0.5
Gradient accumulation steps	4	4
Learning rate scheduler	cosine	cosine
Batch size	1 or 2 (depending on GPU VRAM)	1
Training epochs	2	1
Low-rank adapter parameters (in decoder)	$r = 8, \alpha = 16$, dropout=0.1	$r = 8, \alpha = 16$, dropout=0.1
Low-rank adapter target modules (in decoder)	down_proj, gate_proj, up_proj, k_proj, q_proj, v_proj, o_proj	qkv_proj, o_proj, down_proj, gate_up_proj

Cephalo-8b series chat template

```
User:{prompt_1}<end_of_utterance>
Assistant:{response_1}<end_of_utterance>
User:{prompt_2}<end_of_utterance>
Assistant:{response_2}<end_of_utterance>
...
```

The 4b models use this prompt template:

Cephalo-4b series chat template

```
<|user|>
<|image_1|><|image_2|>...
{prompt_1}<|end|>
<|assistant|>
{response_1}<|end|>
<|user|>
{prompt_2}<|end|>
<|assistant|>
{response_2}<|end|>
...
```

4.5 Protein structure visualizations

Protein structures were visualized using PyMol <https://pymol.org/>.

4.6 Text-to-image generation

We use Stable Diffusion XL Turbo (<https://huggingface.co/stabilityai/sdxl-turbo>) to create 2D images from text prompts generated by Cephalo-Phi-3-vision-128k-4b-beta.

4.7 Text-to-3D model generation

We use Meshy <https://www.meshy.ai/about> to turn a text prompt generated by Cephalo-Phi-3-vision-128k-4b-beta into a 3D model. We use the ‘Text to 3D’ function in Meshy, first create a draft model and then refine in the highest possible quality. The model renders a 3D model with texture information.

4.8 Developing a deeper Cephalo model through model merging

Model merging involves combining components from different models to create a new, hybrid model. This method leverages the strengths of each contributing model, often resulting in improved performance and new capabilities not present in the individual models.

4.8.1 Merging approach and implementation

We created the Cephalo-Idefics-2-vision-10b-alpha model by merging layers from two distinct pre-trained models: lamm-mit/Cephalo-Idefics-2-vision-8b-beta and HuggingFaceM4/idefics2-8b-chatty. The process involved the following steps.

First, we load the two models into memory. The domain-specific fine-tuned model, `lamm-mit/Cephalo-Idefics-2-vision-8b-beta`, is referred to as `model_1`, while the general chat/instruction fine-tuned model, `HuggingFaceM4/idefics2-8b-chatty`, is referred to as `model_2`.

```
from transformers import Idefics2ForConditionalGeneration

model_1_id = 'lamm-mit/Cephalo-Idefics-2-vision-8b-beta'
model_2_id = 'HuggingFaceM4/idefics2-8b-chatty'

vtype = torch.bfloat16 #or other type
model_1 = Idefics2ForConditionalGeneration.from_pretrained(model_1_id, torch_dtype=vtype)
model_2 = Idefics2ForConditionalGeneration.from_pretrained(model_2_id, torch_dtype=vtype)
```

One can use a variety of merging strategies. Here, we use `model_1` in its entirety and add a set of layers from `model_2` to it. For the `lamm-mit/Cephalo-Idefics-2-vision-10b-alpha` model we merge the last $N_m = 8$ layers of the second model. Since the model has $N_t = 32$ layers, layers 24 to 31 are chosen (as discussed in the main text, later layers are used due to their strong contextual understanding and generalization capabilities). Note, we explored this approach for various combinations and have seen generally excellent performance. For instance, we also developed a 12b model by setting $N_m = 16$, that is, the new model consists of the entire Model_1 and half of Model_2, resulting in `lamm-mit/Cephalo-Idefics-2-vision-12b-alpha`.

```
selected_layers_from_model2 = list(model_2.model.text_model.layers)[N_t-N_m:N_t]
```

The selected layers from `model_2` are then concatenated with all layers from `model_1`. This creates a new sequence of layers that integrates the specialized features of `model_1` with the general capabilities of `model_2`.

```
combined_layers = nn.ModuleList(list(model_1.model.text_model.layers) +
                               selected_layers_from_model2)
```

A new model instance (`model_merged`) is created by deep copying `model_1`, and the combined layers are then assigned to this new model:

```
import copy

model_merged = copy.deepcopy(model_1)

model_merged.model.text_model.layers = combined_layers
```

The model is subsequently saved and then fine-tuned. For fine-tuning, we freeze all layers except for the selected layers from Model_2:

```
for name, param in model.named_parameters():
    # Freeze all parameters
    param.requires_grad = False

# Unfreeze the last N_m layers of the text_model
for i in range(-N_m, 0): # Last N_m layers
    layer = model.model.text_model.layers[i]
    for param in layer.parameters():
        param.requires_grad = True

# Optionally, verify the layers that will be trained
for name, param in model.named_parameters():
    print(f"{name}: {param.requires_grad}")
```

A detailed example is provided via <https://github.com/lamm-mit/Cephalo>.

4.8.2 Fine-Tuning

As described in the preceding section, the last N_m layers of the new hybrid model were fine-tuned. This fine-tuning step is crucial as it helped the new model adapt and align the strengths of the merged layers, ensuring that the model

performed well in both domain-specific and general tasks. A learning rate of $2E - 5$ was used, with warm-up and cosine learning rate decay.

4.9 Sparse Mixture of Experts based on the 4b models built on Phi-3-Vision

As another way to create larger, more complex models from smaller ones we employ a sparse mixture of experts (MoE) model to enhance the model. This model leverages multiple expert networks to process different parts of the input, allowing for more efficient and specialized computations. For each token in the input sequence, a gating layer computes scores for all experts and selects the top- k experts based on these scores. We use a softmax (...) activation function to ensure that the weights across the chosen experts sum up to unity.

The output of the gating layer is a set of top- k values (\mathbf{T}) and their corresponding indices (\mathbf{I}). The selected experts' outputs (\mathbf{Y}) are then computed and combined using a weighted sum, where the weights are given by the top- k values. Mathematically, the output for each token (b, s) in the batch is given by:

$$\mathbf{O}_{b,s} = \sum_{i=1}^k \mathbf{T}_{b,s,i} \cdot f_{\mathbf{I}_{b,s,i}}(\mathbf{X}_{b,s})$$

where $\mathbf{T}_{b,s,i}$ is the weight from the gating function for the i -th selected expert, and $f_{\mathbf{I}_{b,s,i}}(\mathbf{X}_{b,s})$ is the output of the i -th selected expert for the token (b, s) . This sparse MoE mechanism allows our model to dynamically allocate computational resources, improving efficiency and performance for complex vision-language tasks. Figure 10 depicts an overview of the architecture.

The code implements the creation of the MoE model directly in custom model code. For instance, as shown via the repository `lamm-mit/Cephalo-Phi-3-MoE-vision-128k-3x4b-beta`, we can easily construct a model consisting of three Phi-3 models:

```
# Initialize the models
expert_models = [model_1, model_2, model_3] # List of expert models
base_model = copy.deepcopy(model_2) # Base model, here chosen to be model_2

# Load a processor (e.g. from base model)
processor = AutoProcessor.from_pretrained(model_name_2, trust_remote_code=True)

# Create the config
config = AutoConfig.from_pretrained(model_name_2, trust_remote_code=True)

# Create the MoE model
moe_config = Phi3VForCausallLMMoEConfig(config=config,
    k=1, num_expert_models=len(expert_models))
moe_model = Phi3VForCausallLMMoE(moe_config, base_model, expert_models, layer_dtype = dtype).to(device)
```

4.9.1 Construction mechanics of the MoE model

To transform an existing model into a Mixture of Experts (MoE) model, we first take the base model use a set of fine-tuned or otherwise trained models to create multiple expert models. Typically, each of the expert models specializes in different aspects of the input data, allowing for greater flexibility and efficiency in processing. To implement this, the original layers of the base model are replaced with modified layers that incorporate the gating and expert mechanisms. A custom configuration class is created to extend the base configuration, adding parameters specific to the MoE setup, such as the number of experts and the number of experts to select in each forward call (k).

Within the algorithm, the original MLP layers in the base model are replaced with a new MoE layer that combines the outputs of the selected experts. This MoE layer uses the gate scores to select the relevant experts' outputs and combines them into a single output by computing a weighted sum. The modified layers are then integrated back into the model, creating a hybrid architecture that retains the original model's structure but with enhanced capabilities. The code is provided at <https://github.com/lamm-mit/Cephalo-Phi-3-MoE>, providing further details.

4.9.2 Training the gating network through sample prompts

The gating networks play a crucial role in determining the appropriate experts for each input token. These gating networks are trained using sample prompts that represent a diverse set of inputs. For each prompt (text only, or

interleaved text-image prompts), we compute the hidden states in each layer and use them to train the gating layer to produce appropriate expert selection scores. The training process involves optimizing the gating layer to minimize the cross-entropy loss between the predicted expert distribution and the target distribution, which encourages the gating network to learn to select the most relevant experts for a given input. The gating layers are trained using collected hidden states from a set of prompts, optimizing the gate values to ensure the effective selection and weighting of expert outputs.

We define a dataset (defined as a Python dictionary of text-image pairs):

```
prompts_per_expert = [
    [{"text": "<|user|>\n<|image_1|>\nPrompt 1 for expert 1<|end|>\n<|assistant|>\n",
     "image": [image_1]},
     {"text": "<|user|>\n<|image_1|>\nPrompt 2 for expert 1<|end|>\n<|assistant|>\n",
     "image": [image_1]}],
    [{"text": "<|user|>\n<|image_1|>\nPrompt 1 for expert 2<|end|>\n<|assistant|>\n",
     "image": [image_2]},
     {"text": "<|user|>\n<|image_1|>\nPrompt 2 for expert 2<|end|>\n<|assistant|>\n",
     "image": [image_2]}],
    [{"text": "<|user|>\n<|image_1|>\nPrompt 1 for expert 3<|end|>\n<|assistant|>\n",
     "image": [image_3]},
     {"text": "<|user|>\n<|image_1|>\nPrompt 2 for expert 3<|end|>\n<|assistant|>\n",
     "image": [image_3]}],
    ...
]
```

The gating network can be trained as follows:

```
# Train gating layers using the provided prompts
gating_layer_params = moe_model.train_gating_layer_params_from_hidden_states(
    processor,
    prompts_per_expert,
    epochs=1000,
    loss_steps=100,
    lr=5e-5,
)

# Set parameters in MoE model
moe_model.set_gating_layer_params(gating_layer_params)
```

While this training function offers good performance, further training may be beneficial especially for deployment of multiple active experts.

Code and data availability

Trained weights, datasets, and other resources for the models utilized in this study can be found at <https://huggingface.co/lamm-mit/cephalo>. Links to various components of this model development are provided at this Hugging Face collection. Additional codes and tools are provided at <https://github.com/lamm-mit/Cephalo> and <https://github.com/lamm-mit/Cephalo-Phi-3-MoE>.

Author contributions

MJB designed the research, conducted dataset generation, the model development and inference. MJB carried out all experiments, simulations and data analysis. MJB wrote and edited the paper.

Conflicts of interest and disclosures

There are no conflicts to declare. Generative AI was used in the creation of this work.

Acknowledgements

We acknowledge support from Google, the Army Research Office, the Office of Naval Research, and the National Institutes of Health (NIH).

References

- [1] K. Guo, Z. Yang, C. Yu, and M. Buehler, Materials Horizons [https://doi](https://doi.org/10.1039/C9MH00039A) (2021).
- [2] F. E. Bock, R. C. Aydin, C. J. Cyron, N. Huber, S. R. Kalidindi, and B. Klusemann, Frontiers in Materials **6** (2019), ISSN 22968016.
- [3] D. C. Elton, Z. Boukouvalas, M. D. Fuge, and P. W. Chung, Molecular Systems Design & Engineering **4**, 828 (2019), ISSN 2058-9689, URL <https://pubs.rsc.org/en/content/articlehtml/2019/me/c9me00039a><https://pubs.rsc.org/en/content/articlelanding/2019/me/c9me00039a>.
- [4] M. Popova, O. Isayev, and A. Tropsha, Science Advances (2018), ISSN 23752548.
- [5] M. J. Buehler, *Accelerating scientific discovery with generative knowledge extraction, graph-based representation, and multimodal intelligent graph reasoning* (2024), 2403.11996.
- [6] W. Lu, D. L. Kaplan, and M. J. Buehler, Advanced Functional Materials p. 2311324 (2023), ISSN 1616-3028, URL <https://onlinelibrary.wiley.com/doi/full/10.1002/adfm.202311324><https://onlinelibrary.wiley.com/doi/abs/10.1002/adfm.202311324>.
- [7] G. X. Gu, C. T. Chen, D. J. Richmond, and M. J. Buehler, Materials Horizons **5**, 939 (2018).
- [8] N. A. Lee, S. C. Shen, and M. J. Buehler, Matter **5**, 3597 (2022), ISSN 25902385, URL <http://www.cell.com/article/S2590238522005902/fulltext>[http://www.cell.com/matter/abstract/S2590-2385\(22\)00590-2](http://www.cell.com/article/S2590238522005902/abstract).
- [9] K. He, X. Zhang, S. Ren, and J. Sun, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society, 2016), vol. 2016-December, pp. 770–778, ISBN 9781467388504, ISSN 10636919.
- [10] K. He, X. Zhang, S. Ren, and J. Sun, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2016-December**, 770 (2015), ISSN 10636919, URL <https://arxiv.org/abs/1512.03385v1>.
- [11] A. Ghafarollahi and M. J. Buehler, *Protagents: Protein discovery via large language model multi-agent collaborations combining physics and machine learning* (2024), 2402.04268.
- [12] B. Ni and M. J. Buehler, Extreme Mechanics Letters **67**, 102131 (2024), URL <https://www.sciencedirect.com/science/article/abs/pii/S2352431624000117>.
- [13] E. Stach, B. DeCost, A. G. Kusne, J. Hattrick-Simpers, K. A. Brown, K. G. Reyes, J. Schrier, S. Billinge, T. Buonassisi, I. Foster, et al., Matter **4**, 2702 (2021), URL <https://www.sciencedirect.com/science/article/abs/pii/S2590238521003876>.
- [14] M. Abolhasani and E. Kumacheva, Nature Synthesis **2**, 483 (2023), URL <https://www.nature.com/articles/s44160-022-00123-2>.
- [15] Z. Qin and M. Buehler, Molecular Simulation **38**, 695 (2012), URL <http://www.tandfonline.com/doi/abs/10.1080/08927022.2012.685943>.
- [16] Q. Zhang, X. Yang, P. Li, G. Huang, S. Feng, C. Shen, B. Han, X. Zhang, F. Jin, F. Xu, et al., Progress in Materials Science **74**, 332 (2015), ISSN 0079-6425.
- [17] Y. Wu, D. U. Shah, C. Liu, Z. Yu, J. Liu, X. Ren, M. J. Rowland, C. Abell, M. H. Ramage, and O. A. Scherman, Proceedings of the National Academy of Sciences **114**, 201705380 (2017).
- [18] L. S. Dimas, G. H. Bratzel, I. Eylon, and M. J. Buehler, Advanced Functional Materials **23**, 4629 (2013), ISSN 1616301X.
- [19] M. Milazzo, N. Contessi Negrini, S. Scialla, B. Marelli, S. Farè, S. Danti, and M. Buehler, Advanced Functional Materials **29** (2019), ISSN 16163028.
- [20] A. R. Studart, Advanced Functional Materials **23**, 4423 (2013), ISSN 1616301X.

- [21] K. Qiu, U. G. K. Wegst, K. Qiu, and U. G. K. Wegst, Advanced Functional Materials **32**, 2105635 (2022), ISSN 1616-3028, URL <https://onlinelibrary.wiley.com/doi/full/10.1002/adfm.202105635><https://onlinelibrary.wiley.com/doi/abs/10.1002/adfm.202105635>
- [22] S. E. Arevalo and M. J. Buehler, MRS Bulletin 2023 pp. 1–14 (2023), ISSN 1938-1425, URL <https://link.springer.com/article/10.1557/s43577-023-00610-8>.
- [23] Y. Hu and M. J. Buehler, APL Machine Learning **1**, 010901 (2023), ISSN 2770-9019, URL <https://aip.scitation.org/doi/abs/10.1063/5.0134317>.
- [24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention is All you Need* (2017), URL <https://papers.nips.cc/paper/7181-attention-is-all-you-need>.
- [25] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. (2023), URL <https://arxiv.org/abs/2307.09288v2>.
- [26] OpenAI (2023), URL <http://arxiv.org/abs/2303.08774>.
- [27] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. (2022), URL <https://arxiv.org/abs/2204.02311v3>.
- [28] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. I. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. (2023), URL <https://arxiv.org/abs/2310.06825v1>.
- [29] S. Gunasekar, Y. Zhang, J. Aneja, C. César, T. Mendes, A. D. Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. De, et al. (2023), URL <https://arxiv.org/abs/2306.11644v2>.
- [30] I. Stewart and M. Buehler, ChemRxiv (2024).
- [31] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, *A comprehensive overview of large language models* (2024), 2307.06435.
- [32] Y. Li, S. Jiang, B. Hu, L. Wang, W. Zhong, W. Luo, L. Ma, and M. Zhang, arXiv preprint arXiv:2405.11273 (2024).
- [33] K. Carolan, L. Fennelly, and A. F. Smeaton (2024), URL <https://arxiv.org/abs/2404.01322v1>.
- [34] E. L. Buehler and M. J. Buehler (2024), URL <https://arxiv.org/abs/2402.07148v1>.
- [35] G. Developers, *Paligemma* (2024), accessed: 2024-05-27, URL <https://ai.google.dev/gemma/docs/paligemma>.
- [36] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. (2023), URL <https://arxiv.org/abs/2303.12712v1>.
- [37] M. J. Buehler, Appl. Mech. Rev. (2023), URL <https://doi.org/10.1115/1.4063843>.
- [38] M. Nejjar, . Luca, Z. †1, F. Stiehle, and I. Weber (2023), URL <https://arxiv.org/abs/2311.16733v3>.
- [39] R. K. Luu and M. J. Buehler, Adv. Science. (2023), URL <https://doi.org/10.1002/advs.202306724>.
- [40] Y. Ge, W. Hua, K. Mei, J. Ji, J. Tan, S. Xu, Z. Li, and Y. Zhang (2023), URL <http://arxiv.org/abs/2304.04370>.
- [41] B. Bhattacharjee, A. Trivedi, M. Muraoka, M. Ramasubramanian, T. Udagawa, I. Gurung, R. Zhang, B. Dandala, R. Ramachandran, M. Maskey, et al. (2024), URL <https://arxiv.org/abs/2405.10725v2>.
- [42] M. Huh, B. Cheung, T. Wang, and P. Isola (2024), URL <https://arxiv.org/abs/2405.07987v1>.
- [43] G. Gemini Team, *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context* (2024), URL <https://goo.gle/GeminiV1-5>.
- [44] A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, et al., Transformer Circuits Thread (2024), URL <https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html>.
- [45] M. J. Buehler, Journal of the Mechanics and Physics of Solids p. 105454 (2023), ISSN 0022-5096, URL <https://linkinghub.elsevier.com/retrieve/pii/S0022509623002582>.
- [46] D. Spivak, T. Giesa, E. Wood, and M. Buehler, PLoS ONE **6** (2011), ISSN 19326203.
- [47] M. Abdin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, et al., *Phi-3 technical report: A highly capable language model locally on your phone* (2024), 2404.14219.
- [48] M. Dehghani, B. Mustafa, J. Djolonga, J. Heek, M. Minderer, M. Caron, A. Steiner, J. Puigcerver, R. Geirhos, I. Alabdulmohsin, et al. (2023), ISSN 10495258, URL <https://arxiv.org/abs/2307.06304v1>.

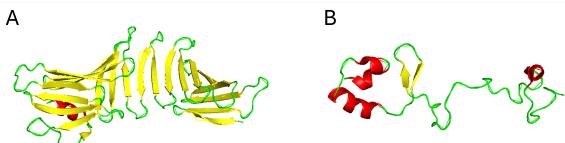
- [49] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, et al., arXiv preprint arXiv:2107.14795 (2021), URL <https://arxiv.org/abs/2107.14795>.
- [50] D. Sen and M. Buehler, *Scientific Reports* **1** (2011), ISSN 20452322.
- [51] C. Goddard, S. Siriwardhana, M. Ehghaghi, L. Meyers, V. Karpukhin, B. Benedict, M. McQuade, and J. Solawetz, arXiv preprint arXiv:2403.13257 (2024).
- [52] T. Akiba, M. Shing, Y. Tang, Q. Sun, and D. Ha, *Evolutionary optimization of model merging recipes* (2024), 2403.13187.
- [53] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen (2021), URL <https://arxiv.org/abs/2106.09685v2>.
- [54] D. Eigen, M. Ranzato, and I. Sutskever, in *International Conference on Learning Representations (ICLR)* (2014).
- [55] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, in *International Conference on Learning Representations (ICLR)* (2017).
- [56] M. I. Jordan and R. A. Jacobs, *Neural computation* **6**, 181 (1994).
- [57] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, et al. (2024), URL <https://arxiv.org/abs/2401.04088v1>.
- [58] S. F. Fischer, M. Thielen, R. R. Loprang, R. Seidel, C. Fleck, T. Speck, and A. Bührig-Polaczek, *Advanced Engineering Materials* **12**, B658 (2010), ISSN 1527-2648, URL <https://onlinelibrary.wiley.com/doi/full/10.1002/adem.201080065>
<https://onlinelibrary.wiley.com/doi/10.1002/adem.201080065>.

Supplementary Materials

Chat between a User and Cephalo-Idefics-2-vision-12b-alpha

User: You are a materials scientist and you carefully analyze images, step by step.

Answer briefly.



User: You are shown two proteins, colored by their secondary structure.

Looking at the protein images, and analyzing the structure, which protein has a more organized structure?

Cephalo: The protein on the left has a more organized structure, as indicated by the presence of alpha-helices and beta-sheets. These secondary structures contribute to the overall stability and function of the protein. The protein on the right, on the other hand, has a more disorganized structure, with fewer or no visible secondary structures. This could imply that the right protein is less stable or functional compared to the left protein.

User: Which protein (A or B) would have a higher mechanical resistance?

Cephalo: Protein A would have a higher mechanical resistance due to its more organized structure, which includes alpha-helices and beta-sheets. These secondary structures provide strength and stability to the protein, making it more resistant to mechanical forces.

User: Suggest a few design improvements.

Cephalo: To improve the design, the following changes could be made:

1. Enhance the mechanical properties by incorporating more hydrogen bonds and disulfide bridges, which would increase the stability and rigidity of the protein structure.
2. Introduce new amino acids or modify existing ones to alter the protein's conformation, potentially improving its function or stability.
3. Optimize the sequence to enhance the folding process, ensuring that the protein adopts a more stable and functional conformation.

User: Be creative and think about using the proteins to generate a new electronic device. How would you go about it?

Cephalo: To create an electronic device using these proteins, the focus would be on leveraging their unique molecular properties to generate specific functions such as conductivity, sensitivity, or responsiveness. This could be achieved by:

1. Functionalizing the proteins with conductive or semiconducting materials to enhance their electrical properties.
2. Integrating the proteins into a composite material, potentially with other biopolymers or nanomaterials, to form a new type of sensor or actuator.
3. Modifying the protein structure to optimize its interaction with other components or materials, allowing for better performance in electronic devices.

By using the proteins' inherent properties, the design would aim to create a new class of materials with unique functionalities that could be applied in various electronic devices, such as flexible sensors, bioelectronics, or biomedical devices.

Text Box S1: Chat between a User and Cephalo-Idefics-2-vision-12b-alpha, focused on assessing two proteins. The model is tasked to figure out structural features and to assess likely mechanical properties based on the images. The two proteins considered are PDB IDs 5B2A (crystal structure of a mutant of OspA, a beta-sheet rich structure with around 250 residues) and 1WVK (NMR solution structure of the partially disordered protein At2g23090 from *Arabidopsis thaliana*).

Chat between a User and Cephalo-Idefics-2-vision-12b-alpha

User: You are a materials scientist and you carefully analyze images, step by step.

Answer concisely, and accurately.



User: You are shown two photos of two very different objects. Carefully study the two photos.

What do you see?

Cephalo: The first image shows a group of ants climbing up a tree. The ants are moving in a line, using their legs to grip onto the surface of the tree. The second image shows a close-up view of a porous material, likely a biological or synthetic material, with a network of interconnected holes. The material appears to be made of a light-colored, porous material, possibly a type of foam or a natural substance. The scale bar in the second image indicates that the structure is on the microscopic scale, with a size of 7 micrometers.

User: Which of the two images would be better suited to create fracture-resistant materials?

Cephalo: The second image, showing the porous structure of the biological or synthetic material, would be better suited for creating fracture-resistant materials. The interconnected holes in the material likely contribute to its strength and ability to withstand fractures, as seen in the ants' ability to climb up the tree without breaking the surface. This natural design can inspire the development of synthetic materials with similar properties.

User: What about the other one, what may it be useful for?

Cephalo: The first image of ants climbing a tree can inspire the design of adhesive materials, as the ants' ability to climb with minimal slippage suggests a strong adhesive force. This natural example can guide the development of synthetic adhesives that mimic the ants' grip on surfaces.

User: How can these concepts be used to inform new materials?

Cephalo: The concepts shown in these images can inspire the design of new materials by drawing from natural examples. The ants' ability to climb a tree without breaking the surface can inspire the development of adhesive materials, while the porous structure of the biological or synthetic material can inspire the creation of fracture-resistant materials. These natural designs can guide the development of synthetic materials with enhanced properties, such as strong adhesion and resistance to fractures.

User: Next, specifically develop a research idea that integrates the concepts seen in the two images.

Cephalo: A research idea could involve the study of the mechanical properties of ants' claws and the porous structure of their climbing surfaces, which might provide insights into the design of new materials with enhanced adhesive and fracture resistance. By understanding the natural mechanisms that allow ants to climb with minimal slippage, we can develop synthetic materials with similar properties. Similarly, by examining the porous structure of natural materials, we can gain insights into the design of new materials that are both strong and lightweight.

User: Provide a step-by-step outlook to future research.

Cephalo: Future research could involve the detailed analysis of the ants' claws and the porous structure of their climbing surfaces, using advanced imaging techniques such as scanning electron microscopy (SEM) and atomic force microscopy (AFM) to understand the mechanical properties at the microscopic level. This could lead to the development of new materials with enhanced adhesive and fracture resistance by mimicking the natural designs observed in the ants and the porous material.

Text Box S2: Chat between a User and Cephalo-Idefics-2-vision-12b-alpha, analyzing an image of ants forming a bridge and a microscopy image of a pollen particle. The model is tasked to reason over both and develop bio-inspired design concepts. Left image, reproduced from https://en.m.wikipedia.org/wiki/File:AntBridge_Crossing_10.jpg, based on a CC BY-SA 4.0, Igor Chuxlancev. Right image, reproduced from https://en.wikipedia.org/wiki/Pollen#/media/File:Arabis_voch1-4.jpg, CC BY-SA 3.0, Marie Majaura.

Chat between a User and Cephalo-Idefics-2-vision-12b-alpha

User: You are a materials scientist and you carefully analyze images, step by step.

Answer concisely, and accurately.



User: Describe what you see. Do you detect a crack?

Cephalo: The image shows a single, uncracked, unbroken, and unadulterated chicken egg placed on a wooden surface. The egg is a natural product, and its surface appears smooth and unblemished. The wooden surface provides a contrasting texture to the egg, highlighting its round shape and smooth shell.

User: Use the image and propose what would happen if I were to drop the object.

Cephalo: If I were to drop the egg, it would likely crack due to the impact. The egg's shell is not designed to withstand significant force, and the impact energy would cause the shell to deform and potentially break. This is because the eggshell is a biomaterial composed of calcium carbonate and protein, which provides some strength and rigidity but is not as resistant to impact as synthetic materials.

Text Box S3: Chat between a User and Cephalo-Idefics-2-vision-12b-alpha, analyzing an image of an egg on a wooden surface. The model correctly states that there is no crack in the egg, and further explains what would happen if the egg is dropped.

Chat between a User and Cephalo-Idefics-2-vision-12b-alpha

User: You are a materials scientist and you carefully analyze images, step by step.

Answer concisely, and accurately.



User: Describe what you see. Do you detect a crack?

Cephalo: The image shows a cracked egg on a wooden surface. The egg is brown and appears to be in the process of breaking, with visible cracks and fissures on its surface. The wooden surface provides a contrasting texture to the smooth, porous surface of the egg. The cracks suggest that the egg has been subjected to some form of stress, likely leading to its breakage.

User: Use the image and propose what would happen if I were to drop the object.

Cephalo: If I were to drop the egg, it would likely shatter into several pieces due to the cracks visible in the image. The egg's brittle nature and the existing damage would cause it to break apart, releasing the contents inside.

Text Box S4: Chat between a User and Cephalo-Idefics-2-vision-12b-alpha, analyzing an image of a pre-cracked egg on a wooden surface. The model correctly identifies the crack in the material and also explains what would happen if the egg is dropped, as requested by the human.