npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-025-01564-y

Fine-tuning large language models for domain adaptation: exploration of training strategies, scaling, model merging and synergistic capabilities

A list of authors and their affiliations appears at the end of the paper

The advancement of Large Language Models (LLMs) for domain applications in fields such as materials science and engineering depends on the development of fine-tuning strategies that adapt models for specialized, technical capabilities. In this work, we explore the effects of Continued Pretraining (CPT), Supervised Fine-Tuning (SFT), and various preference-based optimization approaches, including Direct Preference Optimization (DPO) and Odds Ratio Preference Optimization (ORPO), on fine-tuned LLM performance. Our analysis shows how these strategies influence model outcomes and reveals that the merging of multiple fine-tuned models can lead to the emergence of capabilities that surpass the individual contributions of the parent models. We find that model merging is not merely a process of aggregation, but a transformative method that can drive substantial advancements in model capabilities characterized by highly nonlinear interactions between model parameters, resulting in new functionalities that neither parent model could achieve alone, leading to improved performance in domain-specific assessments. We study critical factors that influence the success of model merging, such as the diversity between parent models and the fine-tuning techniques employed. The insights underscore the potential of strategic model merging to unlock novel capabilities in LLMs, offering an effective tool for advancing AI systems to meet complex challenges. Experiments with different model architectures are presented, including the Llama 3.1 8B and Mistral 7B family of models, where similar behaviors are observed. Exploring whether the results hold also for much smaller models, we use a tiny LLM with 1.7 billion parameters and show that very small LLMs do not necessarily feature emergent capabilities under model merging, suggesting that model scaling may be a key component. In open-ended yet consistent chat conversations between a human and AI models, our assessment reveals detailed insights into how different model variants perform, and shows that the smallest model achieves a high intelligence score across key criteria including reasoning depth, creativity, clarity, and quantitative precision. Other experiments include the development of image generation prompts that seek to reason over disparate biological material design concepts, to create new microstructures, architectural concepts, and urban design based on biological materials-inspired construction principles. We conclude with a series of questions about scaling and emergence that could be addressed in future research.

e-mail: mbuehler@mit.edu

The development of Large Language Models (LLMs)[1–6] has enhanced our abilities for natural language processing (NLP) in scientific and engineering applications, due to significant advancements across diverse domains, from general-purpose applications to specialized fields such as materials science and engineering[7–14]. These models, including prominent open-source architectures like Llama and Mistral, have demonstrated strong capabilities in understanding and generating human-like text. However, their application in technical fields requires fine-tuning strategies that adapt these models to specific domain challenges and technical requirements, which are often poorly understood. In the field of biomateriomics, for instance, researchers aim to develop systematic explorations of knowledge across scales, domains, and application areas, including biological material design inspiration[7,15–19]. These and other challenges can be addressed synergistically using multimodal reasoning engines that have, at their core, capabilities derived from LLMs. One rationale is that LLMs have shown strong capabilities to integrate diverse concepts and provide an integrative modeling strategy for diverse contexts seen in biological materials engineering[7,20].

Fine-tuning LLMs for domain-specific applications involves more than simply retraining on specialized data; it requires the exploration of strategies to endow the model with new knowledge while retaining capabilities learned in earlier training stages, to yield optimal model performance. This is particularly challenging since most of the time it is not feasible to train models from scratch due to cost, or because the original datasets are not available. This is a particular concern in open-source models like Llama or Mistral, where certain details about the training process have been released, but the full datasets during pre-training, fine-tuning, and alignment phases are unknown[4,6].

An often-effective strategy that has been used in earlier work is low-rank adaptation (LoRA)[21], where a set of small trainable low-rank tensors is added to linear layers of a larger model to adapt it towards new capabilities[14,21,22]. While this can be an effective method, there are limits to how much a model can be improved and how much new knowledge can be incorporated. Other research suggested that continued pre-training (CPT) on domain-specific corpora can help better introduce new knowledge within the target domain[23], enhancing its relevance and accuracy. However, this typically requires a host of additional strategies to make a model useful for downstream applications, such as instruction following, chat interaction, agentic use, tool calling, and others. Supervised fine-tuning (SFT) is a method used to refine these models by directly teaching them to perform well on specific tasks through curated datasets. However, the potential for further enhancing model performance and unlocking new capabilities through advanced optimization techniques remains a critical area of exploration. In this context, preference-based optimization strategies, such as Direct Preference Optimization (DPO)[24] and Odds Ratio Preference Optimization (ORPO)[25], have emerged as promising approaches. Unlike traditional Reinforcement Learning (RL) methods[26], which often require explicit reward functions and complex environment models, DPO and ORPO focus on optimizing models based on direct feedback or preferences. These methods offer a flexible and efficient means of refining model behavior, particularly when the goal is to align the model's outputs with human expectations or domain-specific criteria, such as being able to reason over or logically deduce answers in a particular domain, such as materials science.

Another area of recent interest is the practice of model merging[27–29], where multiple, differently trained models are combined to create a new model with potentially superior capabilities. Earlier experiments have shown that this process is not simply additive; as we will show, it leads to highly nonlinear interactions between the parameters of the merged models, resulting in the emergence of new functionalities that neither parent model possessed individually. Such emergent behavior suggests that model merging could be a powerful tool for advancing LLM capabilities, enabling the development of models that are not only more accurate but also more adaptable to complex, real-world challenges.

As this brief review shows, there are a myriad of possible strategies, but relatively little data is available in terms of systematic explorations. LLMs are highly complex models, and training is expensive and time-consuming, and often developers focus on a particular approach that yielded acceptable results. Here we take a different approach and specifically investigate the effects of various fine-tuning and optimization strategies on LLM performance, with a particular focus on a systematic, consistent set of experiments as summarized in Table 1. Figure 1 shows an overview of the training corpus and information processing, covering both the general process of utilizing different types of data (raw, processed/distilled, conversational, etc.) and a visualization of the transformation from individual pieces of information to a structured network of interconnected insights.

The plan of the paper is as follows. We first present a brief introduction into key concepts, then present results along with a discussion, followed by conclusions. We present detailed methods and references to codes and model weights, and discuss our dataset creation, development, and training mechanisms.

Results

We follow the process depicted in Fig. 2 in developing models and conducting assessments. Figure 2A shows a conventional linear training pipeline where a base model undergoes Continued Pre-Training (CPT), followed by Supervised Fine-Tuning (SFT), and then optimized using methods like Direct Preference Optimization (DPO) or Odds Ratio Preference Optimization (ORPO) to produce a trained model. Figure 2B shows an alternative pipeline where, after CPT, SFT, and optimization (e.g., DPO, ORPO), the model is further enhanced by merging it with another fine-tuned model (e.g., a general-purpose model). We note that model merging can be done with models extracted from various training stages, such as after CPT, SFT or at the final stage. The implementation process is structured sequentially, with each step building upon the previous one to enhance the

Table 1 | Summary of model development approaches explored with datasets used, and a brief description of capabilities

Approach	Definition	Example	Dataset type
CPT (Continued Pre-Training)	Continuing the training of a language model on a specific domain or additional data after initial pre-training (no template/instruction format use).	Enhancing model knowledge in specialized fields like materiomics, bioinformatics, or broader fields like materials science.	Raw text from papers plus text with reasoning, summarization, distillation
SFT (Supervised Fine-Tuning)	Fine-tuning a pre-trained model using labeled data with supervised learning techniques.	Adapting models to specific tasks such as QA, reasoning, scientific method, or dialog systems.	Question-answer or instruction-answer pairs, conversations
DPO (Direct Preference Optimization)	Optimizing a model by directly learning from preferences, often involving human feedback or ranking.	Aligning models with human preferences for content generation, physics-awareness, reasoning, or human considerations.	Positive/negative examples
ORPO (Odds Ratio Preference Optimization)	A reference model-free monolithic odds ratio preference optimization algorithm that eliminates the need for a separate preference alignment phase.	Refining models without a baseline reference, using direct odds ratio for preference tasks.	Positive/negative examples
Model Merging	Combining multiple models or checkpoints into a single model, e.g., using techniques like spherical linear interpolation (SLERP).	Creating hybrid models that incorporate strengths from different pre-trained models.	No training data used

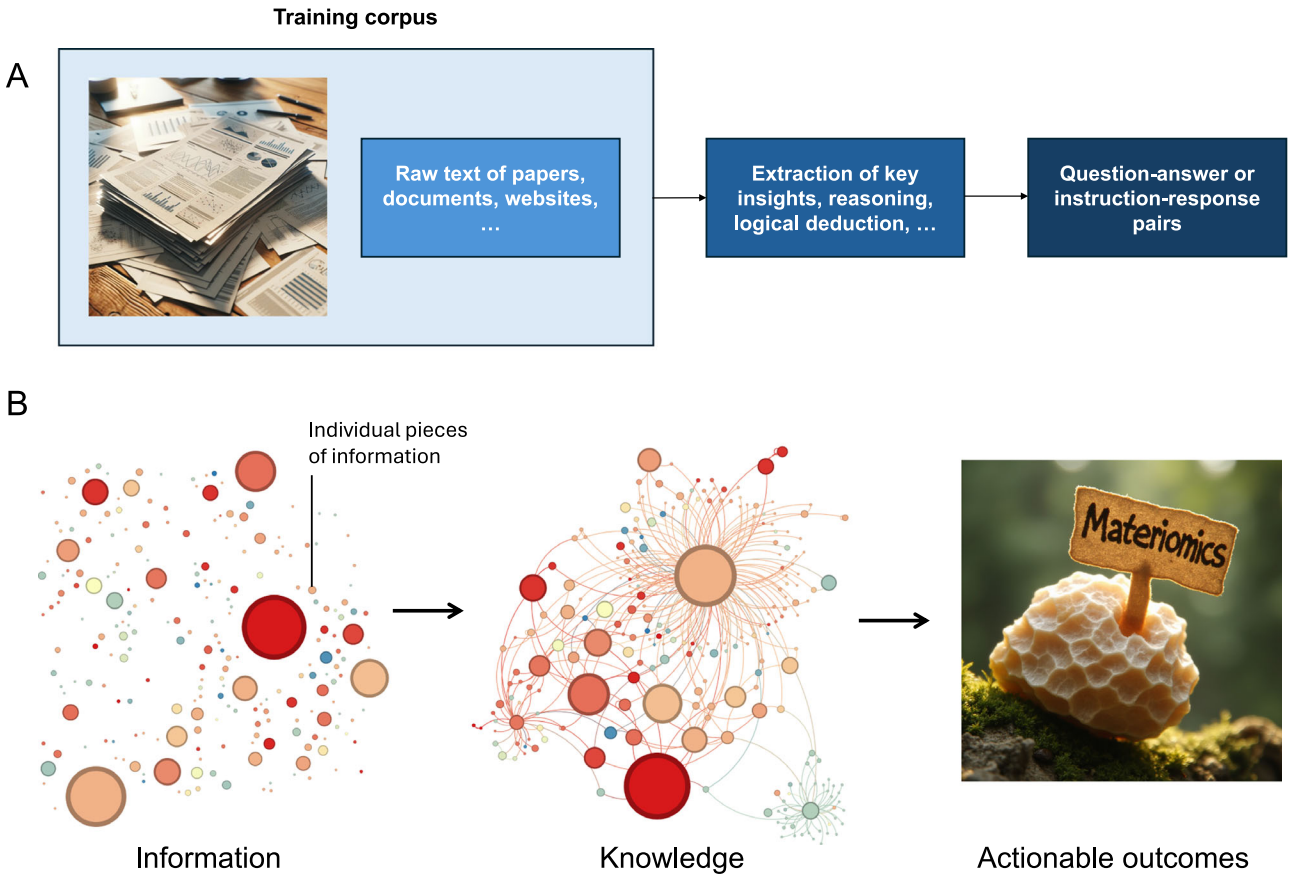


Fig. 1 | Overview of the approach used in this study, including the scientific training corpus and information processing. A The training corpus comprises raw text from various sources such as papers, documents, and websites. This text undergoes extraction of key insights, reasoning, and logical deduction, leading to the generation of question-answer or instruction-response pairs. **B** Visualization of the transformation from individual pieces of information (here shown as scattered nodes of varying sizes) to a structured network of interconnected insights, illustrating the consolidation of knowledge through the training process. This overall schematic illustrates the goals of this research, to build models for complex problems that integrate distinct features, modalities, and concepts. The image above "Actionable outcomes" was generated using lamm-mit/leaf-flux.

model's capabilities in stages. As summarized in Table 1, CPT exposes the model to domain-relevant data, while SFT delivers task-specific instruction using labeled datasets in formats such as question-answer or instruction-response. Preference optimization methods (e.g., DPO or ORPO) align the model with human or baseline preferences established in prior steps. Finally, model merging integrates the strengths of different training paths, leading to improved generalization and, in many cases, the emergence of new capabilities absent from individual models. This structured progression ensures a balance between task-specific accuracy and user preference alignment by addressing knowledge acquisition and preference tuning at distinct stages.

For the purpose of the analysis, we go into the details of model merging strategies. In this work, we focus on Spherical Linear Interpolation (SLERP, details see "Materials and methods" section), as we found it to be the most effective method. SLERP is a mathematical technique originally introduced in the field of computer graphics for smoothly interpolating between rotations represented by quaternions[30]. SLERP has found widespread application in various domains that require smooth transitions between orientations or states, including robotics, physics simulations, as well as real-time graphics. For instance, in robotics, SLERP is used for the practical parameterization of rotations, allowing for seamless motion planning and control[31]. In physics simulations and computer graphics, SLERP is crucial for visualizing and animating rotations in a way that preserves the continuity and smoothness of motion[32,33]. By maintaining the geometric relationships between interpolated states, SLERP ensures that transitions are both smooth and physically meaningful, making it a useful tool in scenarios where precise and continuous interpolation is required. Figure 3 shows the basic concepts of SLERP (versus linear interpolation, LERP), visually. A key aspect of this strategy is that the smooth, nonlinear path helps to preserve the underlying structure of the model parameters. The sphere in this context represents the inherent structure of the model's parameter space, and by maintaining the geometric relationship between the parameters, SLERP ensures that the interpolation respects this original structure and does not puncture it (as linear combination of points would), leading to a more meaningful and coherent blending of capabilities rather than random, unstructured changes. Because the merged points are both congruent with the model geometry (that is, they lie on the sphere used here for demonstration) and because they realize new points previously not accessed, emergent features and capabilities could potentially be unlocked. The smoothness and spherical symmetry assumed in SLERP help preserve angular relationships between parameters, avoiding high-loss regions typically encountered in linear interpolation. This capability is especially beneficial for materials science applications, where parameter space discontinuities and asymmetries are common. SLERP enables better generalization and the development of emergent capabilities, making it a powerful tool in this domain.

In the following, we present a series of results from assessment experiments conducted with different model families and training/merging strategies (details on training, models, datasets, and assessment benchmarks, see "Materials and methods" section). Figure 4 depicts a series of performance evaluations of `Llama-3.1` Model variants across

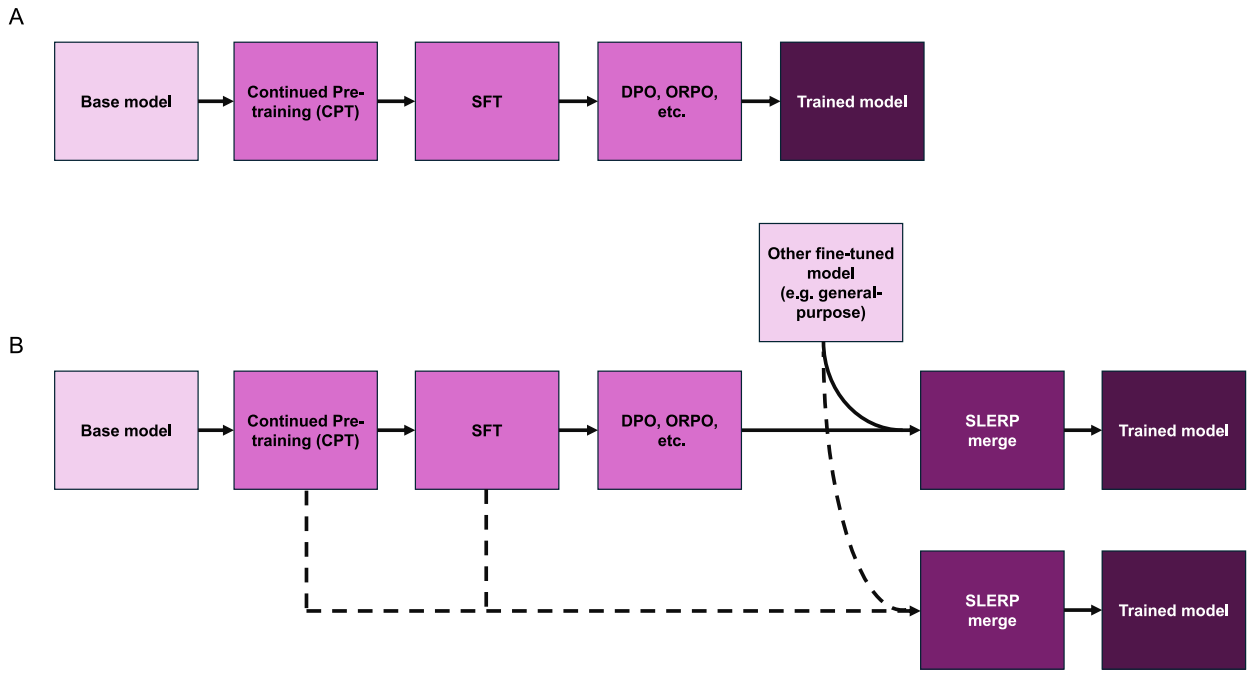


Fig. 2 | Model training, merging and assessment stages. A A conventional training pipeline where a base model undergoes Continued Pre-Training (CPT), followed by Supervised Fine-Tuning (SFT), and then optimized using methods like Direct Preference Optimization (DPO) or Odds Ratio Preference Optimization (ORPO) to produce a trained model. Assessment of the model can be performed at each of the stages, such as using the SFT results for benchmarking. **B** An alternative pipeline where, after CPT, SFT, and optimization (e.g., DPO, ORPO), the model is further enhanced by merging it with another fine-tuned model (e.g., a general-purpose model). Merging can be done with models extracted from various training stages, such as after CPT, SFT or at the final stage.

benchmarks. We use two basic models as the foundation for our training. First, `meta-llama/Meta-Llama-3.1-8B`, the base model of the Llama family that has not been fine-tuned and aligned. Second, the `meta-llama/Meta-Llama-3.1-8B-Instruct` model that has been fine-tuned and aligned to conduct question-answer interactions, along with a host of other capabilities[6]. Except for the LoRA case[14], all of our experiments include CPT (see Table 1 for an overview of the training stages and acronyms used) as the first step, with the aim to endow the base model with domain knowledge from our materials science corpus of papers and distilled, extracted and processed data sourced from scientific studies. We then implement a range of variations, such as CPT only, CPT-SFT, CPT-SFT-ORPO and CPT-SFT-DPO. At each stage, we also implement model merging with the `meta-llama/Meta-Llama-3.1-8B-Instruct` model. Overall, the results reveal that the models that have undergone SLERP merging (especially those combined with DPO and ORPO strategies) generally show the highest accuracy across benchmarks. The best strategy without model merging is found to be the Instruct-CPT-SFT-DPO strategy.

We now conduct the same series of experiments using `Mistral-v0.3` model variants[4] across benchmarks. As in the previous set of results, we use the same dataset across all cases, and we present both non-merged cases and merges with the `mistralai/Mistral-7B-Instruct-v0.3` model. Figure 5 depicts an overview of the performance evaluations across benchmarks for this case. As before, the results show that these models that have also undergone SLERP merging generally show the highest accuracy across benchmarks. The best strategy without model merging is found to be the Base-CPT-SFT strategy, albeit the performance of the Instruct-CPT-SFT strategy is very similar.

The CPT stage involves five epochs. To explore the effect of the number of epochs in this phase, we computed the performance of the direct CPT-SLERP merges for the Mistral models from different training epochs. It is noted that the original merges assessed (and SFT, DPO/ORPO training stages) in Fig. 5 were conducted based on CPT results from epoch 5. Figure 6 depicts a comparison of averaged scores across different epochs for both the Base and Instruct models, using the SLERP method. Figure 6A shows an overview of the results, in a similar format as the earlier performance assessments, depicting performance across all models and variants of CPT epochs used. Figure 6B shows the performance of the Instruct model as a function of the number of CPT training epochs, and Fig. 6C illustrates the performance of the Base model. We can see that the Instruct model demonstrates a consistent improvement in performance with each epoch, peaking at the best score by epoch 5, indicating that it benefits significantly from continued training. In contrast, the Base model shows a more fluctuating performance, with its highest score at epoch 1, followed by slight declines and only a minor recovery at epoch 5. This suggests that while the Base model starts strong, it does not consistently improve with additional training, potentially indicating a saturation point. Both models, however, consistently outperform the baseline score set by the original `mistralai/Mistral-7B-Instruct-v0.3` model, underscoring the effectiveness of the SLERP method, and consistent with the earlier results. The more substantial improvement of the Instruct model over the baseline highlights its robustness in instruction-tuned tasks, making it the preferable choice for such applications, particularly when extended fine-tuning is feasible.

Detailed analysis of key factors in model merging

As the results in Figs. 4 and 5 clearly reveal, SLERP appears to significantly improve model performance due to its ability to respect the geometric properties of the parameter space. However, this analysis did not yet reveal whether we have a significant synergistic effect. To examine this, we plot the results differently, comparing the actual measured performance with an expected performance that is computed by simply averaging the scores of the two parent models. To properly define all key variables, the performance of a merged model is defined as $P_{\text{merged};P_1,P_2}$ (measured per the benchmark), while the expected, averaged score $E(P_1, P_2)$ is calculated as the linear

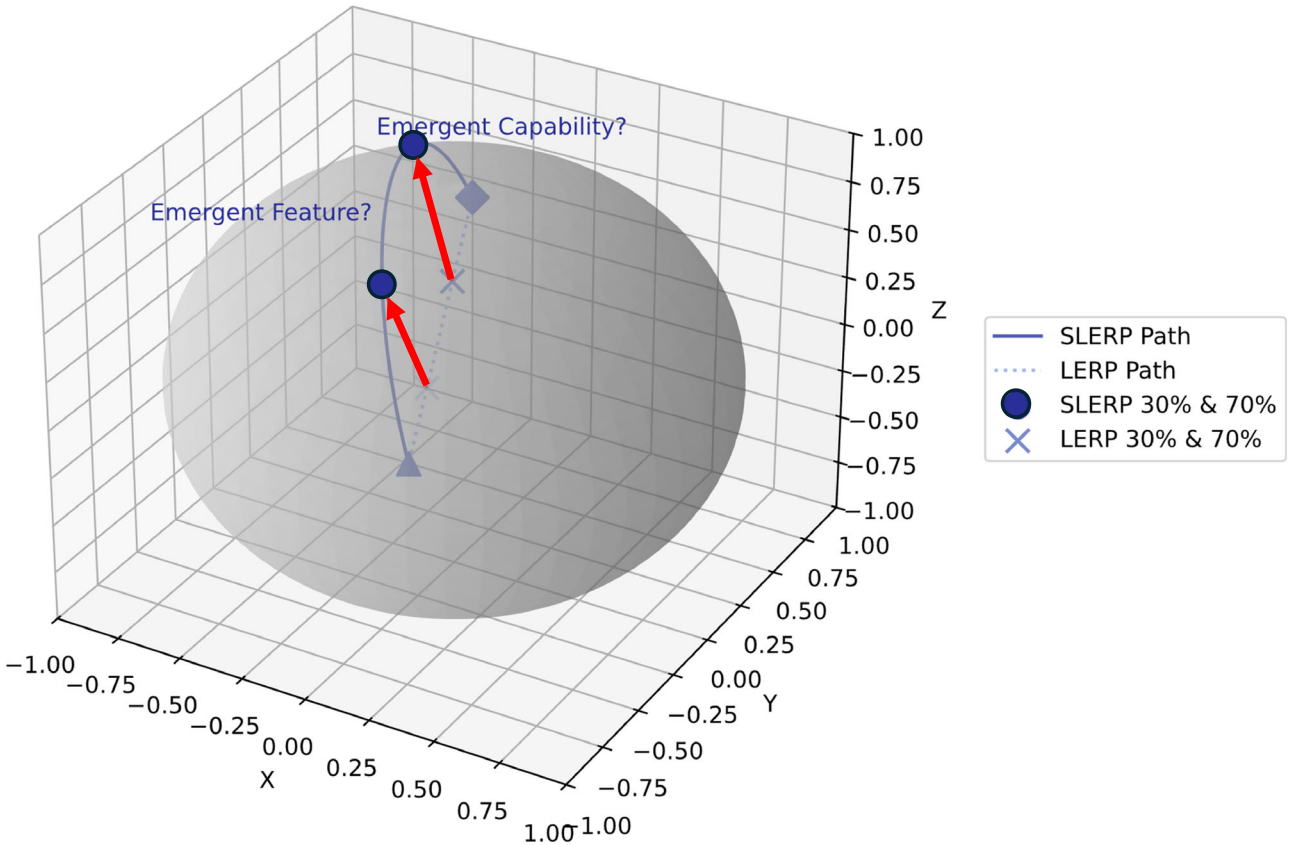


Fig. 3 | Comparison of SLERP (Spherical Linear Interpolation) and LERP (Linear Interpolation) between two points on a unit sphere, illustrating their application in merging Large Language Model (LLM) parameters. SLERP interpolates between points $\mathbf{p}_1$ and $\mathbf{p}_2$ along a spherical path on the surface of the sphere, calculated as SLERP(t), where t is the interpolation parameter (equations see main text). In contrast, LERP interpolates linearly between the same two points, following a straight line through the sphere. Intermediate points at 30% and 70% along both paths are highlighted, showing the difference in how SLERP and LERP handle interpolation. In the context of LLMs, SLERP is particularly effective for merging model parameters from different pre-trained models, facilitating the emergence of new abilities that neither parent model possessed alone. The smooth, nonlinear path of SLERP helps to preserve the underlying structure of the model parameters, represented by the unit sphere, potentially unlocking novel interactions between features that lead to enhanced performance and the development of emergent capabilities. The sphere in this context represents the inherent structure of the model's parameter space, and by maintaining the geometric relationship between the parameters, SLERP ensures that the interpolation respects this original structure and does not puncture it (as the LERP points would), leading to meaningful and coherent blending of capabilities rather than random, unstructured changes. A key point is that because the merged points are both congruent with the model geometry (that is, they lie on the sphere used here for demonstration) and because they realize new points previously not accessed, emergent features and capabilities could potentially be unlocked.

average of the performances of the two parent models:

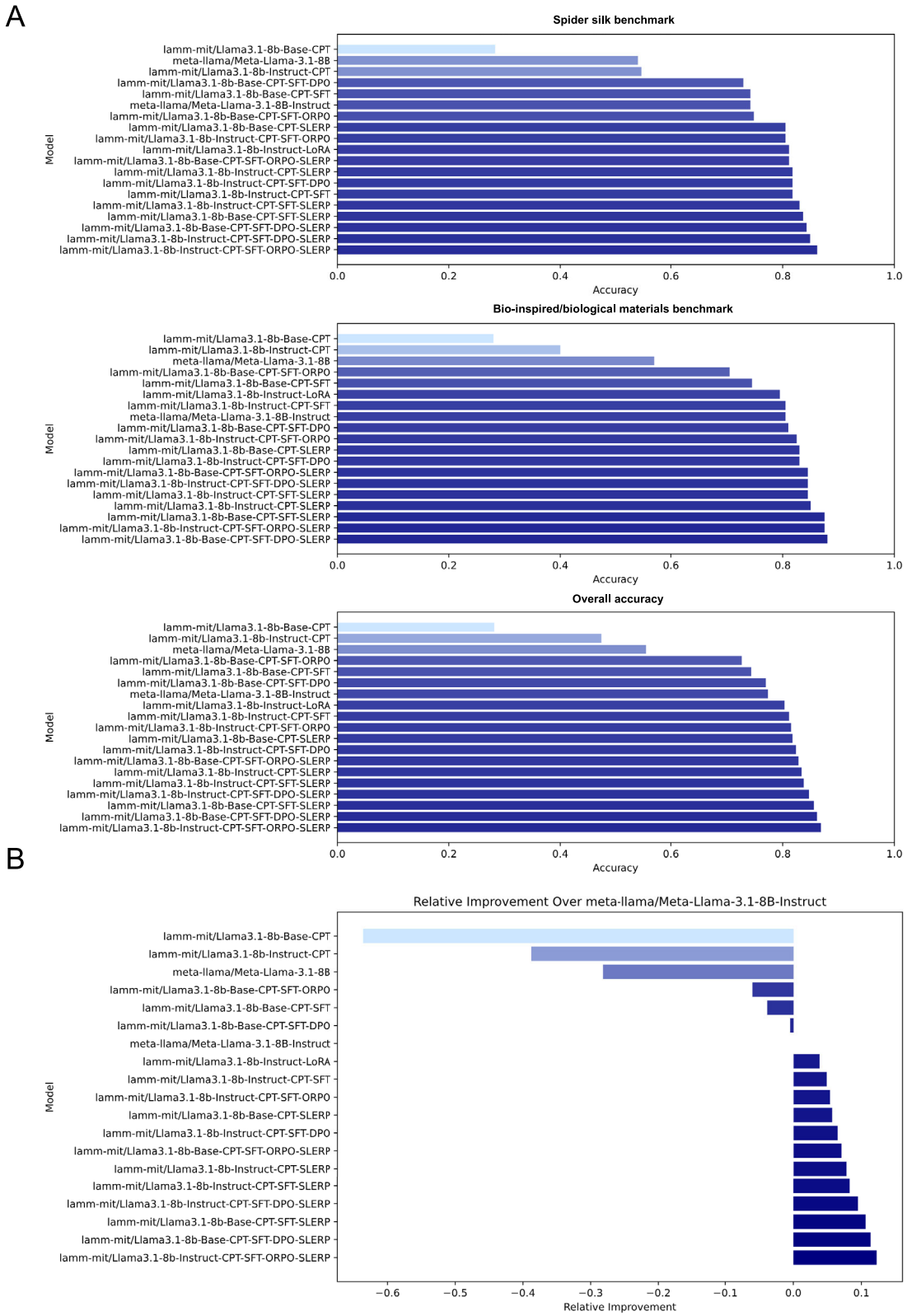
$$E(P_1, P_2) = \frac{P_1 + P_2}{2}.$$

Using these metrics, Fig. 7 shows a detailed exploration of performance of SLERP variants for different cases, plotting the actual observed performance over an estimated, expected score based on a simple average of the score of both parent models (linear combination).

Notably, the strong deviation from the diagonal reveals nonlinear, synergistic effects, where the actual observed model performance is much greater than a simple averaging of the capabilities of the parent models alone. Results are shown for both the `Llama-3.1-8B` and `Mistral-7B-v0.3` model series, respectively, for a variety of training strategies and datasets used in the process. We find that the results are similar for both models. An important distinction that can be seen in the analysis is that for the Llama models, the best-performing model (`lamm-mit/Llama3.1-8b-Instruct-CPT-ORPO-SLERP`) is based off the Llama Instruct model, whereas for the Mistral model (`lamm-mit/mistral-7B-Base-v0.3-CPT-SFT-SLERP`) it is based off the Mistral Base model.

To better understand the mechanics behind the observed effects, we briefly discuss the mathematical underpinnings of SLERP merging. Unlike linear interpolation, which assumes a flat Euclidean space, SLERP explores a richer parameter space by interpolating along a curved path on a unit sphere (we refer also to Fig. 3). This approach allows SLERP to uncover regions in the parameter space that might represent combinations of parameters more effective than those found in either model alone. SLERP further balances the specialized knowledge learned by each model, combining their strengths without simply averaging them. By avoiding high-loss regions that linear interpolation might pass through, SLERP ensures a smoother transition, potentially leading to better generalization in the merged model. The nonlinear nature of SLERP's path also considers the complex interactions between parameters, which can reveal beneficial interactions that a simple linear combination would miss. Furthermore, SLERP may act as a form of regularization, preventing overfitting to the idiosyncrasies of a single model's training data, thus enhancing generalization. Finally, SLERP helps mitigate the effects of catastrophic forgetting, preserving knowledge from both models when one has been fine-tuned or trained after the other. These factors combine to make SLERP a powerful tool for model merging, leading to a merged model that often performs better than either of the original models on their own.

Hence, we believe that the observed effectiveness of SLERP in merging models can be attributed to its ability to enhance nonlinear interactions between parameters by exploring the spherical geometry of



the parameter space. Given two sets of model parameters θ_1 and θ_2, each parameter can be seen as a vector in a high-dimensional space. The interpolation performed by SLERP respects the curvature of this space, allowing for combinations of parameters that are not simply linear but involve deeper, nonlinear synergies (see Fig. 3). Consider the parameters θ_1 and θ_2 as consisting of individual components $\theta_{1,i}$ and $\theta_{2,i}$ in a given layer of the neural network. SLERP combines these parameters as follows:

$$\theta_{i,\text{merged}} = \| \theta_1 \|^{1-t} \| \theta_2 \|^{t} \left(\frac{\sin((1-t)\omega)}{\sin(\omega)} \hat{\theta}_{1,i} + \frac{\sin(t\omega)}{\sin(\omega)} \hat{\theta}_{2,i} \right)$$

Fig. 4 | Performance evaluation *P* of `Llama-3.1` model variants across benchmarks. A Accuracy results for various variants on different benchmarks: Spider Silk, Bio-inspired/Biological Materials, and Overall Accuracy. The models were evaluated after undergoing different training and optimization strategies (Continued Pre-Training (CPT), Supervised Fine-Tuning (SFT), Direct Preference Optimization (DPO)/Odds Ratio Preference Optimization (ORPO), and model merging). **B** Relative improvement of model variants over the `meta-llama/Meta-Llama-3.1-8B-Instruct` baseline model. This highlights how each training strategy contributes to the model's performance gains or losses across the various benchmarks, providing insight into the effectiveness of different approaches. It is notable that models that underwent CPT, SFT, and to some extent preference optimization (e.g., DPO, ORPO) show a deterioration in performance, as indicated by negative relative improvement values. However, after applying the Spherical Linear Interpolation (SLERP) merging technique, these same models exhibit significant performance gains, surpassing the baseline model. This highlights the effectiveness of model merging in combining the strengths of different specialized models, resulting in a robust final model with superior overall performance. Overall, the results show that the models that have undergone SLERP merging (especially those combined with DPO and ORPO strategies) generally show the highest accuracy across benchmarks. Merging in this case is always done with `meta-llama/Meta-Llama-3.1-8B-Instruct`. All models have been trained with the same datasets in all stages, as shown in Table 6.

This combination allows for interactions between $\theta_{1,i}$ and $\theta_{2,i}$ that are nonlinear in nature. For example, if $\theta_{1,i}$ and $\theta_{2,i}$ represent weights connected to different features in the network, their spherical combination could activate a new feature ϕ_i that is not present in either model individually:

$$\phi_i = f\left(\theta_{i,\text{merged}} \cdot x_i\right)$$

where x_i is the input feature and $f(\cdot)$ is the activation function. The nonlinear combination of parameters may lead to new behaviors or capabilities, as the interpolated parameters could synergistically enhance or suppress features in ways that the individual models cannot.

SLERP avoids destructive interference by maintaining the angular relationships between the parameter vectors, which can prevent the loss of specialized features learned by either model. The spherical symmetry imposed by SLERP introduces a regularization effect, smoothing the transition between the models and enabling the merged model to generalize better. This process often results in the emergence of new capabilities or improvements in performance that neither of the original models possessed.

The ability of SLERP to uncover these new capabilities can also be understood through the lens of overparameterization and the principles of ensemble methods. Overparameterized neural networks are known to generalize well, even when trained to zero error, due to their increased capacity to capture complex patterns[34]. SLERP leverages this capacity by combining parameters in a nonlinear fashion, effectively utilizing the high-dimensional space in which these parameters reside. As a result, the merged model can exhibit emergent properties that are not apparent in either of the original models. SLERP's mechanism resembles ensemble methods, where combining diverse models leads to better generalization[35]. In this case, the diversity comes from the different training histories and learned features of the two models. The spherical interpolation pathway created by SLERP acts as a continuum of model ensembles, where at each point along the path, the combined parameters may activate new and beneficial feature interactions. SLERP not only preserves the strengths of the individual models but also has the potential to generate entirely new capabilities through its sophisticated interpolation method. This makes it a useful tool for our goal to merge models that complement each other or to create a more versatile and generalizable model from existing pre-trained models.

We examine the variations and potential trends in the strategies explored using clustering analysis. Figure 8 provides a comprehensive clustering analysis of SLERP strategies applied to both the `Llama-3.1-8b` and `Mistral-7B-v0.3` models, and the resulting impact on their performance. We explore the use of two methods. First, K-Means clustering, a partition-based method that groups data into a predefined number of clusters by minimizing the distance between data points and the cluster centroids, providing insight into the natural groupings of models based on their expected and actual performance. Second, we use hierarchical clustering, an agglomerative method that creates a tree-like structure, a dendrogram, to show the nested relationships between models at various levels of similarity, revealing the hierarchical organization and potential subgroupings within the data.

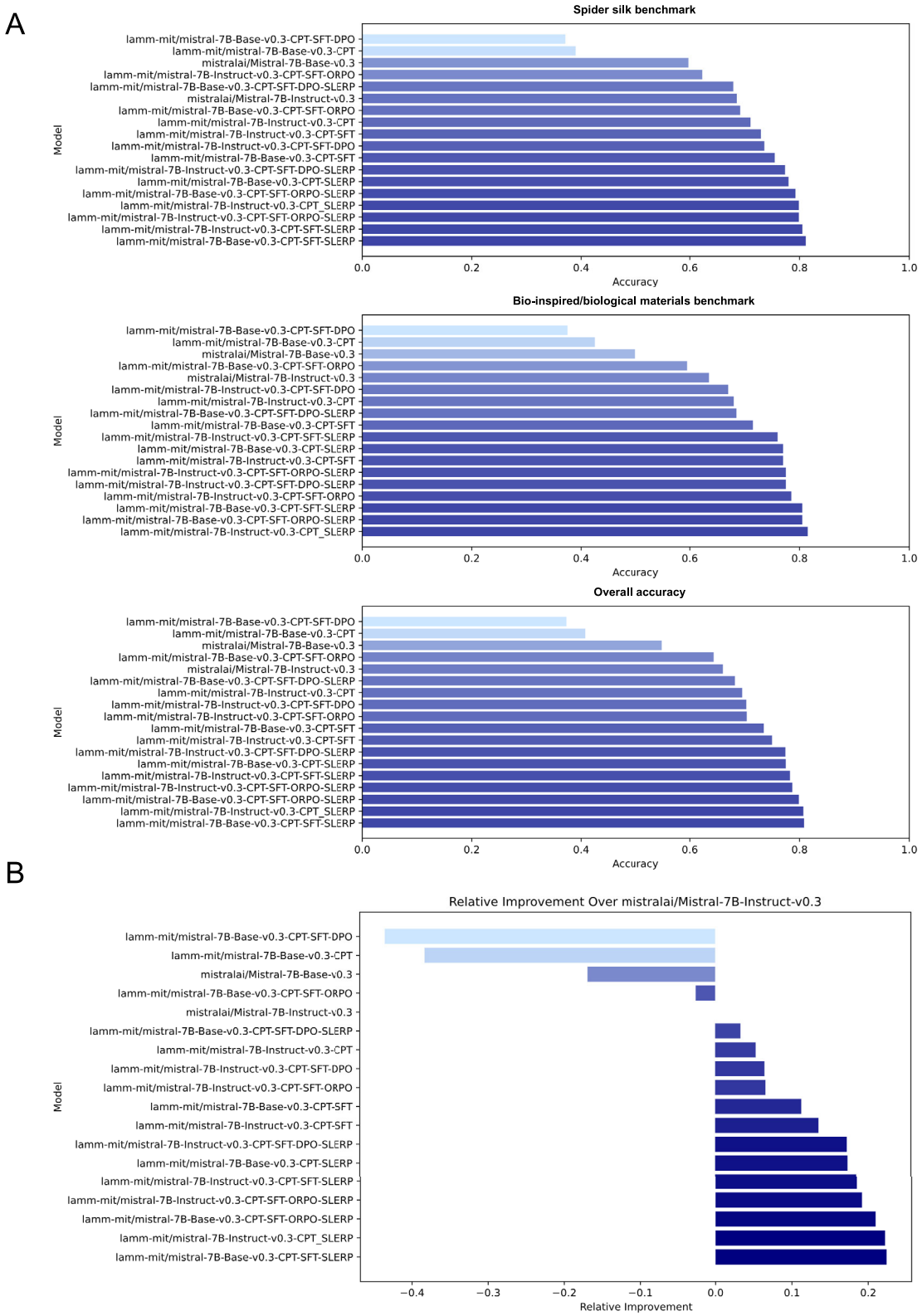
Figure 8 A illustrates the K-Means clustering of the `Llama-3.1-8b` models using standardized expected and actual scores, with Gaussian KDE (Kernel Density Estimation) applied to visualize the centroids. The analysis reveals distinct groupings that correspond to different SLERP strategies, indicating that specific merging techniques produce closely related performance outcomes. Figure 8B presents a similar K-Means clustering for the `Mistral-7B-v0.3` models. Here too, distinct clusters emerge, showing that the SLERP strategies significantly influence the models' performance profiles. Notably, the clustering patterns observed in the Mistral models are more pronounced compared to the Llama models, suggesting that the Mistral architecture might be more sensitive to these optimization and merging strategies.

Across both the Llama and Mistral models, the K-Means analysis clearly delineates two performance-based clusters. Models that incorporate multiple fine-tuning strategies, especially ORPO, consistently form clusters with higher actual scores, outperforming models that rely on simpler strategies. This suggests that the complexity and thoroughness of the fine-tuning process play a crucial role in achieving better model performance, as indicated by the clustering results. Next we explore hierarchical clustering as a way to better break down these distinctions.

To do this, we use a dendrogram analysis. A dendrogram is a tree-like diagram that displays the arrangement of clusters generated through hierarchical clustering. This visualization helps elucidate the relationships among the models, with closely related models (in terms of performance) clustering together. The dendrogram reveals that models employing similar training strategies are grouped into distinct subclusters, highlighting the effectiveness of these approaches in shaping model performance. Figure 8C introduces the hierarchical clustering dendrogram for the `Llama-3.1-8b` models, and Fig. 8D for the `Mistral-7b` models. The dendrogram demonstrates how different models cluster together, indicating similar performance outcomes. When comparing the dendrograms of the Llama and Mistral models, it becomes evident that while both models are positively influenced by SLERP strategies, the Mistral models show more defined clustering patterns. This suggests a stronger impact of the various strategies on the Mistral architecture.

Figure 9 shows the effect of using a larger CPT dataset using the extended dataset of 8000 papers, but with more varied format including more defected text, when training the Llama series. As can be seen, performance decreases, underscoring the effect of higher quality, clean data for positive training outcomes. As mentioned earlier, the extended dataset was constructed using a mix of PDF2Text and Nougat OCR[36]; we found these methods to yield more variable text quality. While, for instance, Nougat can successfully render equations in Markup format, it also leads to a relatively frequent occurrence of unknown symbols, page breaks, repeated characters, and other defects. These methods did not cause issues when the data was further processed into question-answer pairs or summaries, rational explanations, and so on, but there is an apparent negative effect on CPT as the data is provided in raw format.

Likewise, a similar test case with the extended dataset was conducted with the Mistral series of models. The variant trained on the original integrated dataset achieved the best overall benchmark of 0.81, whereas the variant trained on the extended dataset achieved 0.80. These results suggest that future experiments could be conducted to assess the effect of this particular dataset variation on that model architecture's performance. We



leave this to future work; noting that the overall effects of model merging and the use of Base vs. Instruct models as basis are stable, with differences, however, in which exact strategy yields the best results: For the Llama and Mistral models, it was Instruct-CPT-SFT-ORPO-SLERP and Instruct-CPT-ORPO-SLERP. These observations are further complicated by the effect of prompting, which can skew results one way or the other. An overarching theme, however, is that consistently, SLERP merging yields super performance. For a straightforward and computationally effective way to implement a fine-tuning strategy, the procedure Instruct-CPT-SLERP is probably the best overall choice. While it does not yield the best performance for all scenarios, it generally yields strong performance. The differences show that nuanced benchmarking and prompt engineering can be critical.

Fig. 5 | Performance evaluation *P* of `Mistral-7B-v0.3` model variants. **A** Accuracy results for various Mistral-7B-v0.3 model variants on the Spider Silk, Bio-inspired/Biological Materials, and Overall Accuracy benchmarks. Initial models trained with Continued Pre-Training (CPT) and Supervised Fine-Tuning (SFT) show moderate performance. Models further optimized using Odds Ratio Preference Optimization (ORPO) or Direct Preference Optimization (DPO) exhibit significant improvements in accuracy across all benchmarks. Model merging results in further significant improvements. The relative improvements are even more pronounced than those seen in the `Llama-3.1` models (here exceeding 20% versus around 12%), indicating the particular effectiveness of these techniques for the Mistral series. **B** Relative improvement of model variants over the baseline `mistralai/Mistral-7B-Instruct-v0.3` model. The Base model subjected to CPT alone initially shows a decrease in relative performance. However, after SFT, ORPO and especially after applying Spherical Linear Interpolation (SLERP) merging, especially with ORPO or DPO optimization, these models demonstrate substantial positive relative improvement, surpassing the baseline by a greater margin than the improvements seen in the `Llama-3.1` models. This highlights the powerful impact of these combined strategies in enhancing the overall performance of the Mistral models. It is notable that a direct merge of the Base-CPT-SFT model results in significant performance, close to the Instruct-CPT-SFT strategy. Merging is always done with `mistralai/Mistral-7B-Instruct-v0.3`. The same training set is used for all experiments, as defined in Table 6.

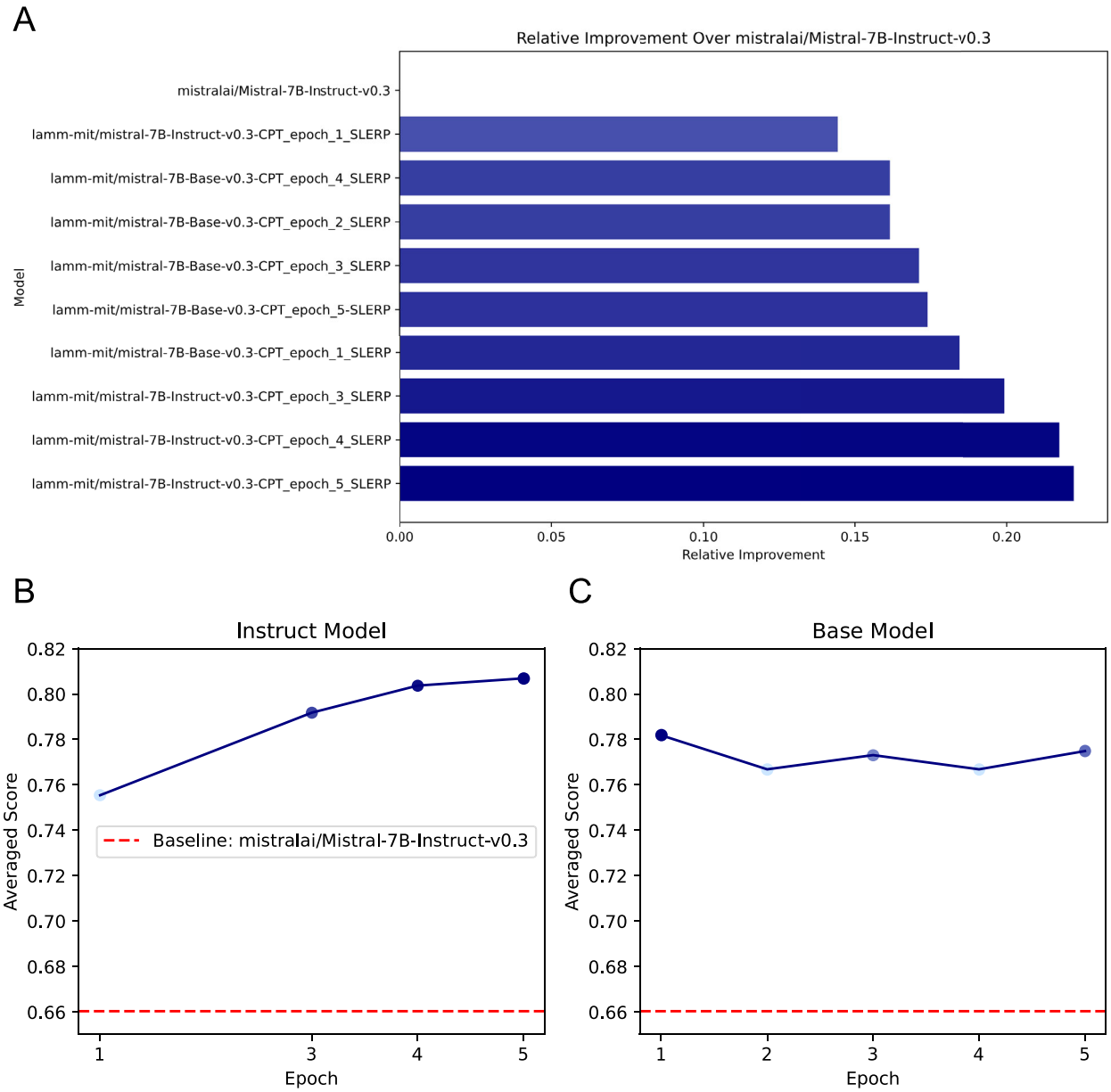


Fig. 6 | Comparison of averaged scores across different epochs for both the Base and Instruct models fine-tuned with the Spherical Linear Interpolation (SLERP) method. The baseline score for the `mistralai/Mistral-7B-Instruct-v0.3` model is indicated by the red dashed line in both subplots. **A** shows an overview of the results, in similar format as the earlier performance assessments (Figs. 4 and 5), showing performance across all models and variants of Continued Pre-Training (CPT) epochs used. **B** shows the performance of the Instruct model over five training epochs, while **C** subplot illustrates the performance of the Base model over five epochs.

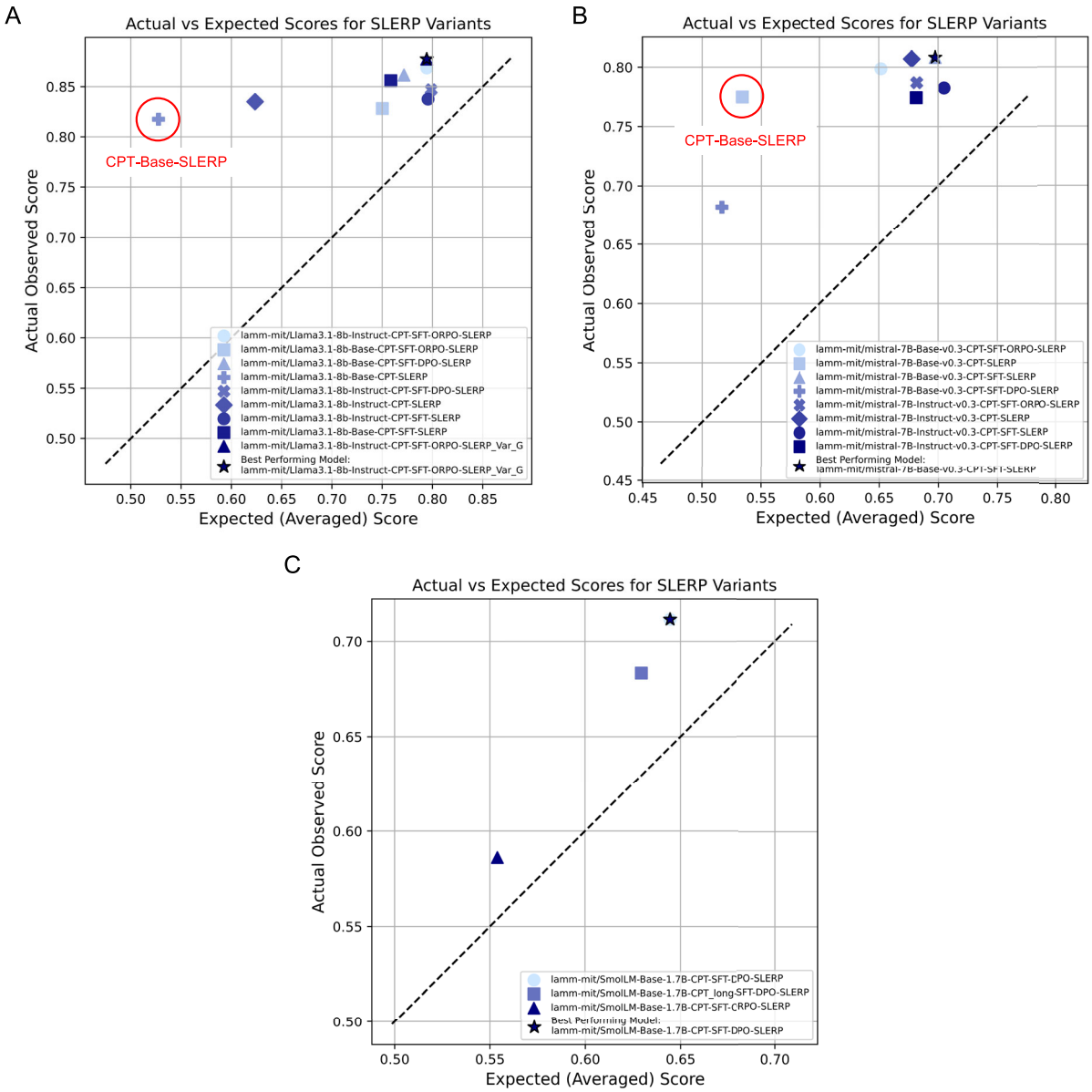


Fig. 7 | Exploration of performance of Spherical Linear Interpolation (SLERP) variants for different cases, plotting the actual observed performance of the merged model $P_{merged;P_1,P_2}$ over a linear, expected average score $E(P_1, P_2)$ based on a simple average of the score of both parent models (linear combination). The deviation from the diagonal shows clear nonlinear, synergistic effects, where the actual observed model performance is much greater than a simple averaging of the capabilities of the parent models alone. Results are shown for both the `Llama-3.1` (**A**) and `Mistral-7B-v0.3` (**B**) model series, respectively. **C** shows results for the much smaller `SmolLM` family of models, where the deviation of the observed score from the expected score is not as significant, and even below the best performance of one of the pre-merged models (see Fig. 13 for a detailed analysis). Results for the Llama and Mistral models are similar across all experiments and show a clear nonlinear effect of model merging. As marked with a red circle, ∘, the CPT-Base-SLERP strategy tends to yield some of the highest deviation from the expected score and is, at the same time, a relatively straightforward training strategy.

Mechanistic analysis to elucidate key steps with highest impact on performance

As a next step in the analysis, we focus on correlation heatmaps to illustrate the relationships between various model attributes and the performance of merged models. As shown in Fig. 10, the performance of a merged model is denoted as P_{merged}, while the performance of the two parent models is denoted as P_1 and P_2. Performance improvement is defined as the difference between the performance of the merged model and the maximum performance of the two parent models:

$$\text{Performance Improvement} = P_{\text{merged}} - \max(P_1, P_2)$$

Diversity between parent models is measured as the absolute difference between their individual performances:

$$\text{Diversity} = |P_1 - P_2|$$

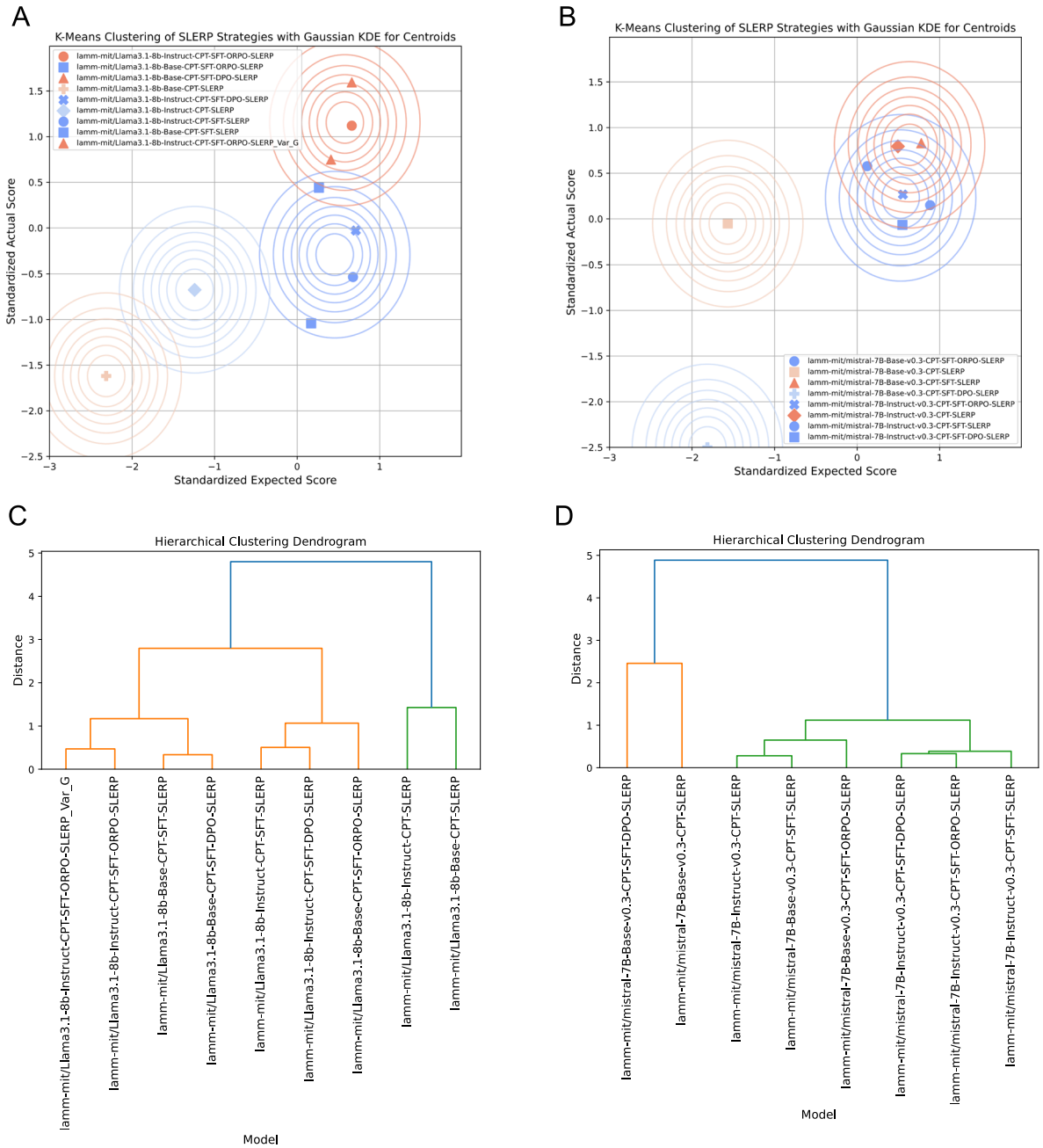


Fig. 8 | Clustering analysis of Spherical Linear Interpolation (SLERP) strategies and hierarchical clustering dendrograms for model performance. A K-Means clustering of `Llama-3.1-8b` models using standardized expected and actual scores, with Gaussian KDE (Kernel Density Estimation) applied to visualize the centroids. The clustering reveals distinct groupings based on the performance outcomes of various SLERP strategies. **B** K-Means clustering of `Mistral-7B-v0.3` models using the same approach as in (**A**). Similar to the Llama models, distinct clusters emerge, highlighting the different performance profiles of the models post-SLERP merging. **A** Hierarchical clustering dendrogram for the `Llama-3.1-8b` models based on the clustering analysis. **B** Hierarchical clustering dendrogram for the `Mistral-7B-v0.3` models. The dendrograms in (**C** and **D**), respectively, reveal how different models cluster together, indicating that these strategies yield similar performance outcomes. The comparison between the `Llama-3.1-8b` (**C**) and `Mistral-7B-v0.3` dendrograms (**D**) shows that while both models respond well to SLERP strategies, the Mistral models exhibit more distinct clustering patterns.

To elucidate overall trends that can be gleaned from the results, Fig. 10 depicts correlation heatmaps for the fine-tuned Llama and Mistral models. The data reveals distinct relationships between various metrics. In Llama models, a strong negative correlation between diversity and SFT suggests that higher diversity reduces reliance on supervised fine-tuning, whereas performance improvement shows moderate positive correlations with both merged performance and SFT, indicating that these factors contribute to improved outcomes. In contrast, Mistral models exhibit a more robust positive correlation between performance improvement and merged performance, especially in instruction-tuned models, where the Base model type significantly enhances merged performance. ORPO, while contributing to performance improvements in both models, has a more pronounced impact in Mistral models. Overall, the findings suggest that diversity tends to reduce SFT dependency, particularly in Llama models, while instruction-tuned Base models in Mistral benefit more from merging strategies, emphasizing the importance of model selection and optimization methods.

In model merging, there are several parameter choices, including the relative density of the parameters that are preserved across the

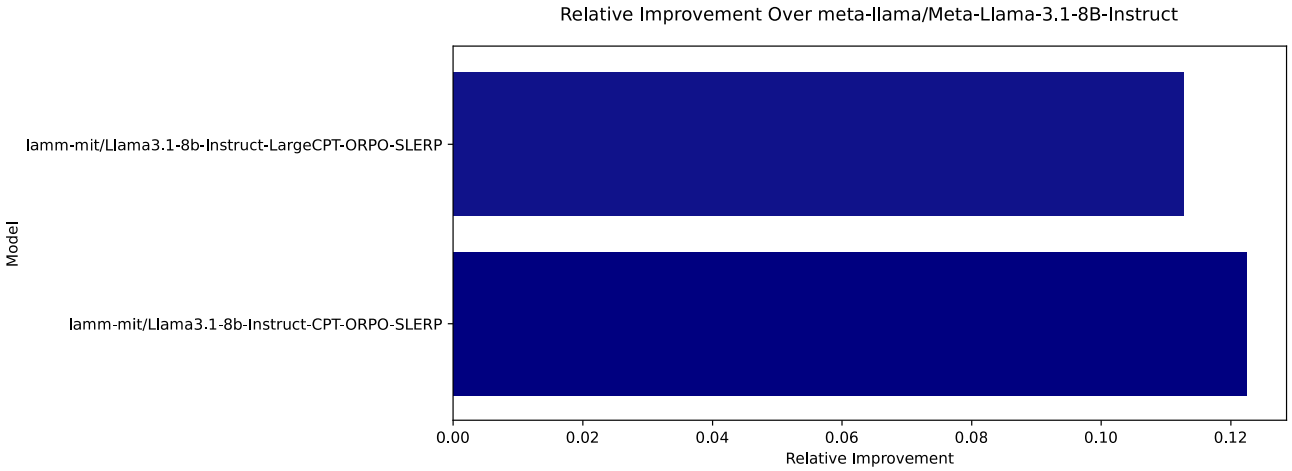


Fig. 9 | Effect of using a larger Continued Pre-Training (CPT) dataset using the extended dataset, along with the `lamm-mit/magpie-ultra-v0.1` dataset, including a more varied quality with a higher content including more defected text. As can be seen, performance decreases, underscoring the effect of higher quality, clean data for positive training outcomes. Future experiments could be done to discern these effects more clearly, especially focusing on the effect of various training data compositions.

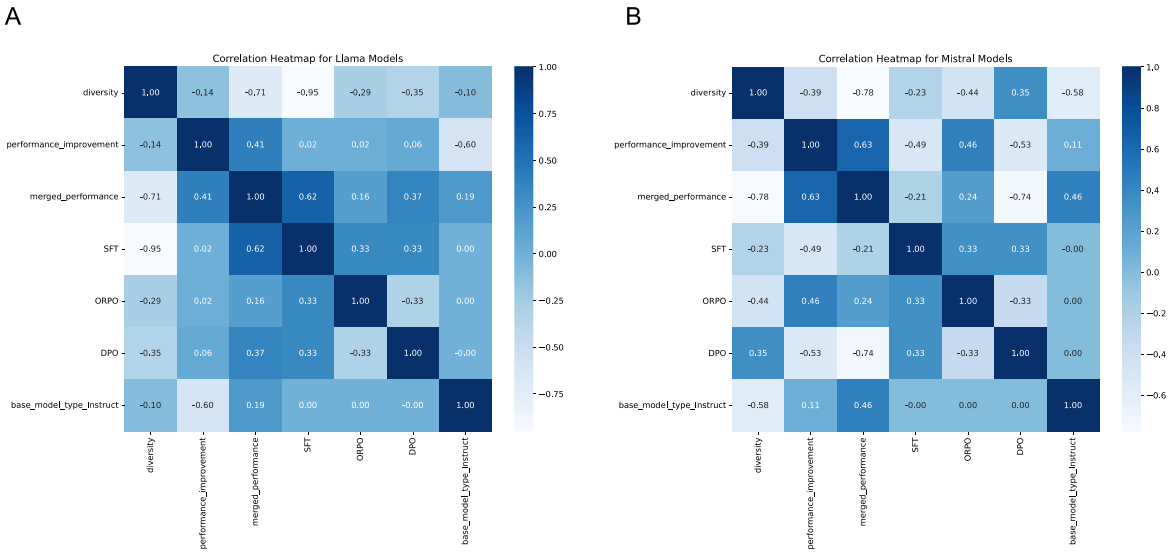


Fig. 10 | Visualization of correlation heatmaps to assess the relationships between various model attributes and the performance of merged models, where the performance of the merged model is the primary outcome of interest. The attributes considered include the diversity between parent models, the performance improvement relative to the parent models, Supervised Fine-Tuning (SFT), Direct Preference Optimization (DPO)/Odds Ratio Preference Optimization (ORPO), and whether the model is based on the Base or Instruct architecture. The correlation coefficients range from −1 to 1, with positive values indicating a direct relationship between the attribute and the merged model performance, and negative values indicating an inverse relationship. **A** Llama Models: SFT shows the strongest positive correlation with merged performance, suggesting that models incorporating SFT tend to achieve better results after merging. Conversely, diversity between parent models has a strong negative correlation with merged performance, implying that greater differences between parent models are associated with lower merged performance. ORPO and the Instruct architecture exhibit moderate positive correlations with merged performance, indicating that these factors also contribute positively, though less significantly than SFT. **B** Mistral Models: In the Mistral models, performance improvement shows a robust positive correlation with merged performance, particularly in instruction-tuned models, which also show a strong positive correlation between the base model type and merged performance. Diversity, however, exhibits a negative correlation with merged performance, similar to the Llama models, though the effect is less pronounced. ORPO demonstrates a moderate positive correlation with performance improvement, suggesting that this optimization method contributes to enhanced performance in Mistral models, albeit not as strongly as SFT in Llama models. The findings suggest that instruction-tuned base models and merging strategies play a crucial role in optimizing Mistral model performance, while diversity and SFT influence these outcomes differently in Llama models.

layers of the LLMs being merged. This is exemplified in Fig. 3 for the two points merged at 30% vs. 70% along their SLERP paths. In Fig. 11 we conduct a systematic analysis on variants of the original SLERP merge used in the earlier examples, with a range of alternative options, for the best-performing strategy in the case of the `Llama-3.1` Model variants (CPT-SFT-ORPO-SLERP). As depicted visually, we vary the self-attention filter values distinctly from the multilayer perceptron (MLP) values (Fig. 11A). Different weighting schemes are employed, starting from the reference case that was chosen based on earlier work[29]. Resulting performance measures are summarized in Fig. 11B, showing that Variants G and F show the best performance (where G is a simple linear progression across the depth of the LLM) (detailed performance assessments for other variants are shown in Fig. 11).

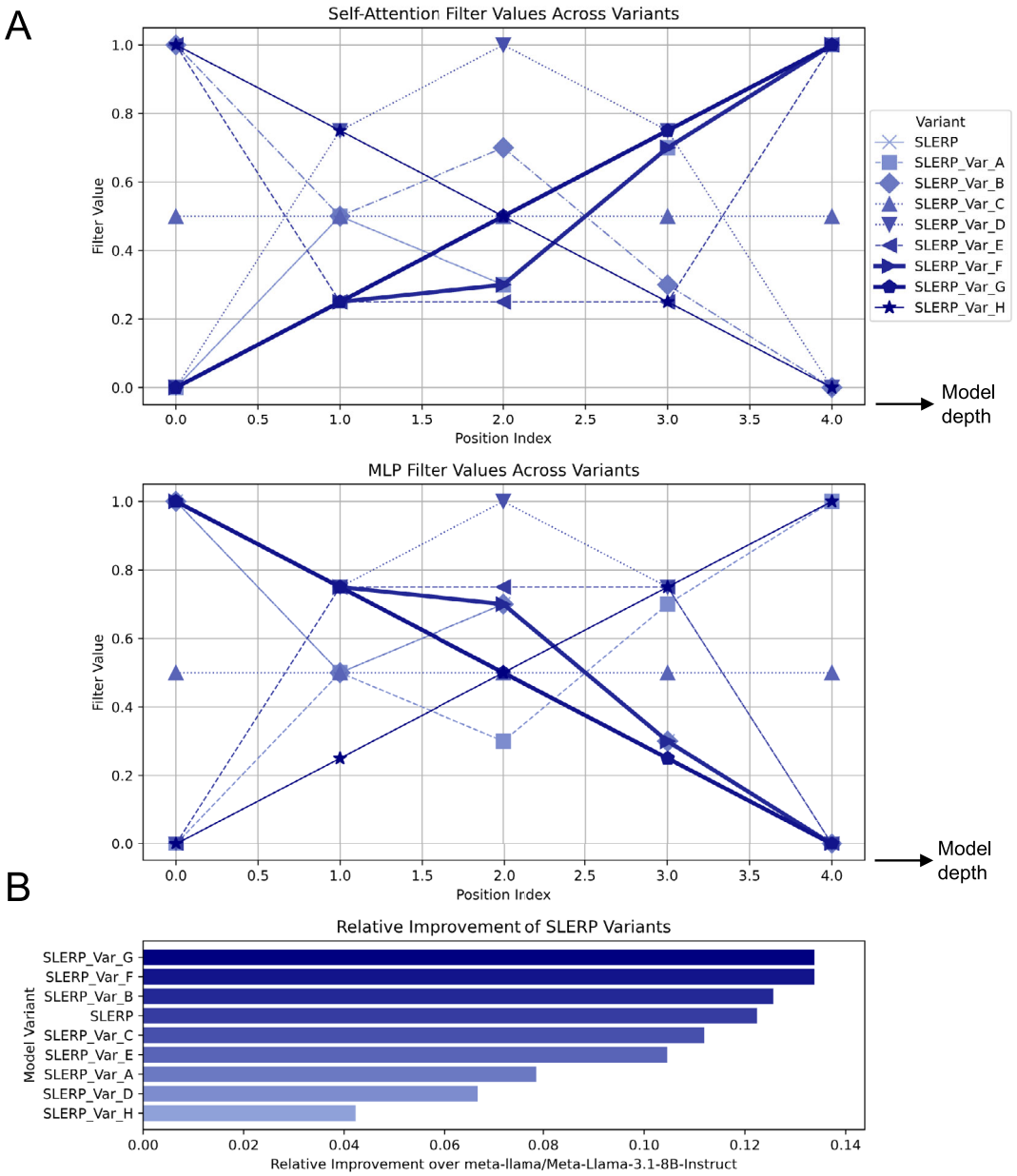
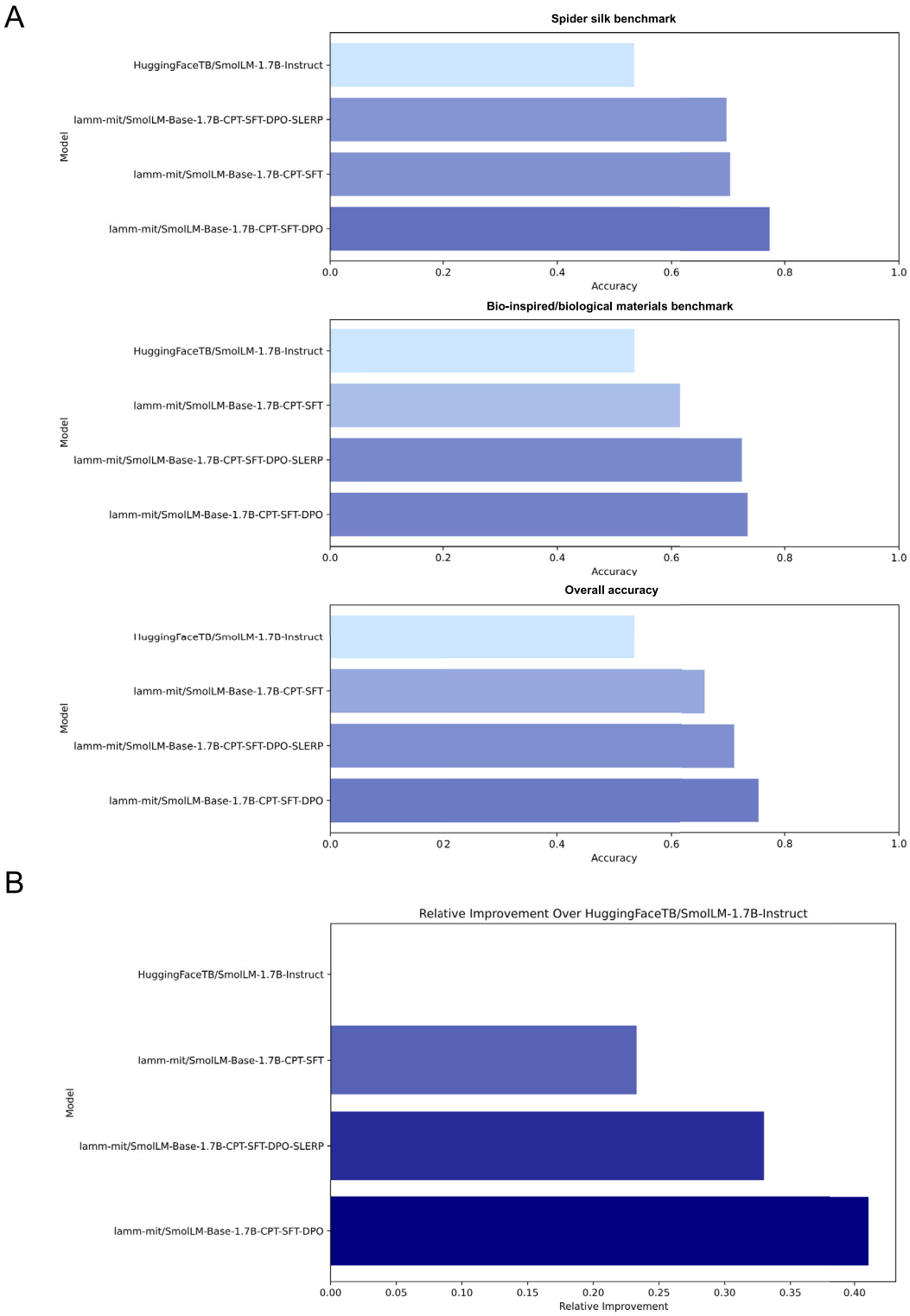


Fig. 11 | Model merging incorporates varying the relative density of how parameters are combined across the layers of the large language models (LLMs) involved (see in Fig. 3 for the two points merged at 30% vs. 70%). The analysis shown here focuses on variants of the original Spherical Linear Interpolation (SLERP) merge used in the earlier examples, for the best-performing strategy in the case of the `Llama-3.1` Model variants(CPT-SFT-ORPO, then merge). As depicted visually, we vary the self-attention filter values distinctly from the multilayer perceptron (MLP) values (**A**). Different weighting schemes are employed, starting from the reference case that was chosen based on earlier work[29]. Resulting performance measures are summarized in (**B**), showing that Variants G and F show the best performance (where G is a simple linear progression).

Contrasting assessments with very small LLMs

While the models studied earlier were modest in size, around 7–8 billion parameters, recent research has resulted in even smaller, yet useful, models that can be particularly useful for edge computing applications, or deployment on devices such as mobile phones or robotic systems. We now examine whether such models also show the marked effects observed earlier due to model merging. We conduct this analysis using the `SmolLM` model series, specifically the 1.7 billion parameter model. This choice is partially motivated by the complete open access of the model, training strategy, and training data. As in the earlier analyses, we start with the base model and successively apply CPT, SFT and DPO (we found that for this small model, DPO worked better than ORPO). Though never reaching the level of absolute performance of fine-tuned 7B or 8B models, in almost all fine-tuning cases with SmolLM, we find the most significant performance increases relative to its original model with the CPT-SFT-DPO version of SmolLM being the top performing variant.

As depicted in Fig. 12 while we observe a significant emergence of new capabilities when applying SLERP to large-scale language models in the 7B and 8B parameter ranges, these emergent behaviors were absent in smaller models, such as those with 1.7B parameters. This may suggest a threshold effect where SLERP's potential to unlock novel abilities is contingent on model size. Smaller models might lack the same level of complexity as larger 7B–8B models that have notably richer high-dimensional parameter spaces and capabilities, especially for reasoning and knowledge recall. These findings



underscore the importance of model scale in the manifestation of emergent properties and provide critical insights into the interplay between interpolation techniques and model complexity. Our results contribute to the broader understanding of scaling laws in neural networks, highlighting the conditions under which advanced capabilities may be realized. A summary of the observed performance over the expected, averaged performance of the base model is shown in Fig. 7C.

Further quantification of the effects of model merging across all model architectures

To better understand whether or not and to what degree model merging improves performance over either one of the two models used for merging, we present the analysis shown in Fig. 13. The plot shows performance deviation of SLERP merged models compared to the best original model used as source.

Fig. 12 | Comparative performance analysis of SmolLM-1.7B models across benchmarks. A Accuracy results for various SmolLM-1.7B model variants on the Spider Silk, Bio-inspired/Biological Materials, and Overall Accuracy benchmarks. The `HuggingFaceTB/SmolLM-1.7B-Instruct` model serves as the baseline, albeit training is done solely on the Base model (`HuggingFaceTB/SmolLM-1.7B`). While applying Continued Pre-Training (CPT) and Supervised Fine-Tuning (SFT) strategies improves performance, the addition of Direct Preference Optimization (DPO) yields further accuracy gains. However, unlike in the much larger Llama or Mistral models, here, Spherical Linear Interpolation (SLERP) merging does not yield the best performance overall. This is more clearly shown in (**B**), where we plot the relative improvement over the baseline model (`HuggingFaceTB/SmolLM-1.7B-Instruct`) across the benchmarks. Notably SLERP combined with DPO yields a slight reduction in performance over the CPT-SFT-DPO case, in stark contrast to the earlier results for Llama and Mistral. The emergent behaviors triggered by SLERP in larger models are not observed here, indicating a potential threshold effect. This suggests that SLERP's ability to unlock novel capabilities may depend on the scale of the model, with smaller models like the 1.7B parameter SmolLM failing to exhibit these emergent properties. These findings underscore the critical role of model scale in realizing advanced capabilities, contributing to the broader understanding of scaling laws in neural networks. Note, in (**B**), the top bar is zero since this model is used as the reference to compute model improvement. It is kept in the visualization for consistency with the earlier plots (where some variants yielded performance degradation, that is, negative values).

The results reveal that the deviation in performance between models merged using SLERP and their best-performing original counterparts, whereby the deviation is calculated as the difference between the best original model's performance and the SLERP model's performance. Hence, negative deviations, where SLERP underperforms relative to the best original model, are marked in red. Positive deviations, indicating better performance of the SLERP model, are shown in shades of blue, with darker blue representing greater improvements.

Interactive examples for open-ended cross-material reasoning and material design tasks

In our next experiment, we conduct interactive conversations with a set of the models, using consistent system prompts and identical user input. We aim to test multi-turn capabilities of the models, assess responsiveness to system prompts and instructions, and the capability to produce structured output (JSON). We will further assess the quality of synthesis of each model, along a set of criteria that include depth of reasoning, creativity, clarity, and whether or not quantitative predictions are featured. Each of the conversations unfolds as follows:

Conversation with: ...

System: Your are a helpful and friendly assistant. You are an expert in materials science with special knowledge in biological materials. You provide concise but accurate responses. You are an expert in combining different concepts, often not obviously related, and provide accurate accounts of correct relationships or analogies.

User: Collagen and leaves, discuss their relationship.

Assistant: ...

User: Thank you! Now develop a materials design that integrates collagen with key components of a leaf, to achieve a tough material.

Assistant: ...

User: Summarize these design features in JSON format.

Assistant: ...

Figures 14–18 present the results of conversations of a human user with a selection of five models (best-performing models, and DPO trained models).

All five models demonstrate a strong ability to connect seemingly unrelated concepts, such as collagen and leaves, and to propose innovative materials designs. The output features rich Markup formatting (note, the raw source is shown here). The models show a consistent understanding of the biological and materials science concepts involved, suggesting that their fine-tuning on domain-specific content has been effective. The differences in the responses mainly pertain to the depth of explanation, the creativity of the proposed designs, and the clarity of the output in both natural language and structured formats like JSON.

As shown in Fig. 14, the model lamm-mit/Llama3.1-8b-Instruct-CPT-SFT-DPO provides a well-organized and detailed discussion, drawing clear parallels between the structure and function of collagen and leaves. The proposed material design is robust, incorporating key components like a collagen-based matrix, vascular-like channels, and mesophyll-like cells. The response is notable for its comprehensive breakdown of each component's role, leading to a thorough and scientifically grounded design. The JSON summary is precise, reflecting the structure of the proposed design effectively.

Figure 15 shows results for the `lamm-mit/Llama3.1-8b-Instruct-CPT-SFT-ORPO-SLERP-Var_G` model. This model delivers a more concise but also insightful analysis of collagen and leaves. The material design focuses on integrating collagen fibrils with cellulose microfibrils and chloroplast-inspired nanoparticles. This resembles an inventive approach to enhancing the material's properties. This model excels in identifying the potential applications of the designed material, showcasing a broader vision for its use. The JSON representation is clear and well-structured, effectively summarizing the design features.

Figure 16 captures results of the `lamm-mit/mistral-7B-v0.3-Base-CPT-SFT-DPO` model. The responses are found to be comprehensive, with a strong focus on the mechanical properties of collagen and leaves. The proposed material design is detailed, incorporating collagen fibrils, cellulose nanofibers, silk fibroin, and a nanocellulose-based matrix. This model particularly stands out for its emphasis on the integration of these components to enhance the material's toughness and durability. The JSON summary is thorough, capturing the complexity of the design and its potential applications.

Next, Fig. 17 shows results for the `lamm-mit/mistral-7B-v0.3-Base-CPT-SFT-SLERP` model. This result illustrates a more straightforward and less detailed analysis compared to the others. While the connection between collagen and leaves is adequately explained, the material design is simpler, focusing on collagen fibers, cellulose nanofibers, chlorophyll, and pectin. This response is notable for its clarity and

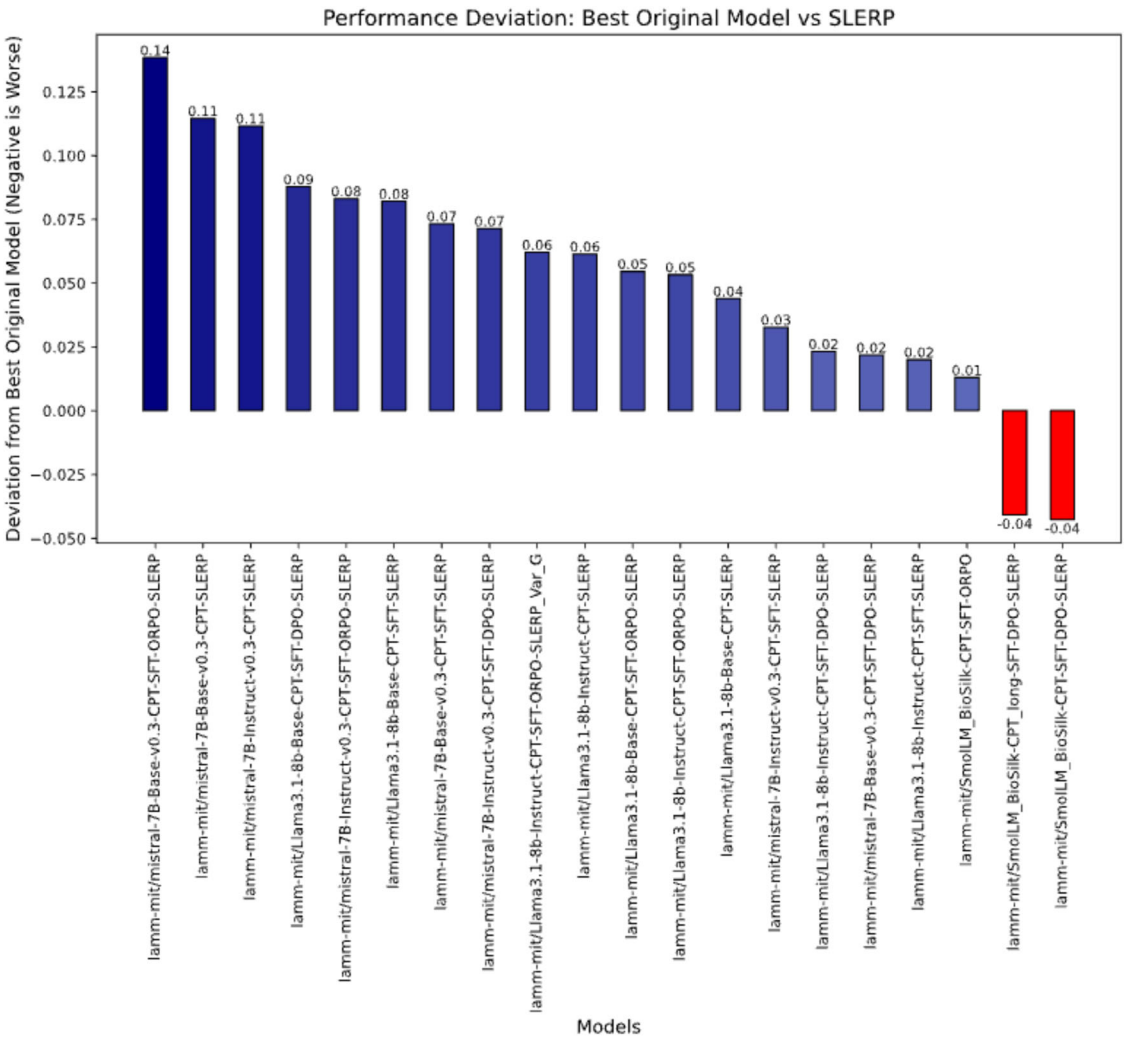


Fig. 13 | Performance deviation of Spherical Linear Interpolation (SLERP) merged models compared to the best original model. The plot illustrates the deviation in performance between models merged using SLERP and their best-performing original counterparts (either without SLERP or with instruction tuning). The deviation is calculated as the difference between the best original model's performance and the SLERP model's performance. Negative deviations, where SLERP underperforms relative to the best original model, are highlighted in red. Positive deviations, indicating better performance of the SLERP model, are shown in shades of blue, with darker blue representing greater improvements. Models are ordered from most significant improvement to most significant underperformance.

simplicity, making it accessible but perhaps lacking the depth seen in other models. The JSON summary is basic but effective in conveying the key elements of the design.

Finally, Fig. 18 showcases the results of the smallest model in this study, `lamm-mit/SmolLM-Base-1.7B-CPT-SFT-DPO`. The model offers an inventive, creative, and highly detailed response, integrating a broad range of components such as collagen fibrils, nanocrystalline cellulose, and an alginate adhesive, resulting in a material creatively referred to as "Leafy-Coraline (LC) Composite". This model excels in proposing a novel composite material with self-healing and shape-memory properties, reflecting a relatively high level of creativity and technical understanding. The JSON summary is comprehensive, capturing the innovative aspects of the design and its potential applications effectively. In spite of its size, this model provides excellent responses.

A summary of key observations is shown in Table 2. We can see that each model demonstrates strengths in different areas, from detailed explanations and innovative designs to clear and concise JSON summaries. The variations in depth, creativity, and technical detail among the models highlight the diversity of approaches and the potential for each to be suited to different types of tasks or applications. Overall, these models provide a strong foundation for further exploration and development in the field of bio-inspired materials design. A detailed breakdown of the analysis, generated using `GPT-4o` by analyzing the raw text of the conversations, is shown in Table 3, showing the results of an analysis conducted by sharing the raw text with `GPT-4o` and prompting the model to identify criteria and assess the different versions. We emphasize that GPT-4o is used as a supporting evaluator, not the definitive authority. LLM-based evaluation minimizes subjectivity and ensures reproducibility across tasks and models compared with human assessment. The potential of LLMs as evaluators has been explored in refs. 37,38 and evaluated in refs. 39–41, and GPT-4o is supported by studies beyond OpenAI's GPT-report[42].

Agentic use in image generation: applications in cross-domain knowledge integration for materials and urban design

We show several examples to highlight the capabilities and potential of fine-tuned LLMs, showcasing a particular application in materials design. The overall goal is to explore how the LLMs developed here can be used to reason over complex materials principles and use the insights developed through multi-step prompting to create a prompt for image generation. We note that our fine-tuned SmolLM based model performs extremely well for this task and yields quite creative prompts that integrate various ideas and concepts (underscoring its potential as a creative agent), and the visual representation of conceptual designs guided by the prompts are generated via a custom

Conversation with: `lamm-mit/Llama3.1-8b-Instruct-CPT-SFT-DPO`

System: Your are a helpful and friendly assistant. You are an expert in materials science with special knowledge in biological materials. You provide concise but accurate responses. You are an expert in combining different concepts, often not obviously related, and provide accurate accounts of correct relationships or analogies.

User: Collagen and leaves, discuss their relationship.

Assistant: At first glance, collagen and leaves may seem like unrelated materials. However, they share some interesting similarities in terms of their structure and function.

Collagen is a protein that forms the structural framework of many tissues in our body, including skin, tendons, ligaments, cartilage, and bone. It has a unique triple-helix structure, known as a tropocollagen helix, which gives it exceptional mechanical properties such as tensile strength and elasticity.

Similarly, leaves have a hierarchical structure that allows them to perform essential functions for plant growth and survival. The main components of a leaf include the epidermis (the outermost layer), mesophyll cells (where photosynthesis occurs), and vascular tissue (which transports water and nutrients). Like collagen, these structures must be strong enough to withstand environmental stresses while also being flexible enough to adapt to changing conditions.

One key similarity between collagen and leaves is their ability to self-assemble into complex networks. In collagen, this process involves the formation of fibrils, which then aggregate into larger fibers and bundles. Similarly, the various cell types and tissues within a leaf work together to form a functional unit that can respond to external stimuli.

Another connection between collagen and leaves lies in their response to stress. Both materials exhibit nonlinear elastic behavior, meaning that their stiffness changes in response to applied force. This property allows them to absorb energy and recover from deformation, which is crucial for maintaining their integrity under varying loads.

In summary, although collagen and leaves come from vastly different sources, they share important structural and functional similarities that allow them to perform similar roles in their respective systems.

User: Thank you! Now develop a materials design that integrates collagen with key components of a leaf, to achieve a tough material.

Assistant: To create a tough material that combines the strengths of collagen and key components of a leaf, we could consider integrating several features:

1. Collagen-based matrix: We start by creating a collagen-based matrix, which provides a robust foundation for the composite material. This matrix would mimic the structure of collagen fibrils found in natural collagen, allowing for high tensile strength and toughness.

2. Vascular-like channels: Next, we incorporate vascular-like channels into the collagen matrix. These channels would mimic the transport system found in leaves, allowing for efficient distribution of nutrients and waste products throughout the material. The channels could be designed using a combination of synthetic polymers and collagen, ensuring both flexibility and durability.

3. Epidermal-like layers: To further enhance the material's performance, we add epidermal-like layers on either side of the collagen matrix. These layers would provide additional protection against environmental stressors and help regulate moisture content, much like the epidermis in leaves. They could be composed of a biocompatible polymer, such as poly(lactic acid) (PLA), which offers good mechanical properties and biodegradability.

4. Mesophyll-like cells: Finally, we introduce mesophyll-like cells into the material. These cells would mimic the photosynthetic capabilities of leaves, providing a source of energy for the material. They could be created using a combination of living cells and synthetic polymers, ensuring both functionality and stability.

By integrating these four components - collagen matrix, vascular-like channels, epidermal-like layers, and mesophyll-like cells - we can create a tough, multifunctional material that combines the best attributes of collagen and key components of a leaf. This material could find applications in biomedical engineering, where its ability to distribute nutrients and support cellular activity would be particularly valuable.

User: Summarize these design features in JSON format.

Assistant:
```
{
    "name": "Collagen-Leaf Composite Material",
    "description": "Tough, multifunctional material combining collagen and key components of a leaf.",
    "components": [
        {
            "type": "matrix",
            "material": "collagen"
        },
        {
            "type": "channels",
            "material": "synthetic polymer + collagen"
        },
        {
            "type": "layers",
            "material": "poly(lactic acid)"
        },
        {
            "type": "cells",
            "material": "living cells + synthetic polymer"
        }
    ],
    "functions": [
        "High tensile strength and toughness",
        "Efficient nutrient and waste product distribution",
        "Environmental stress resistance",
        "Moisture regulation",
        "Photosynthetic capability"
    ]
}
```

Fig. 14 | Conversation generated by the model. `lamm-mit/Llama3.1-8b-Instruct-CPT-SFT-DPO`.

version of a FLUX.1 [dev] model[43] fine-tuned using a image dataset of leaf microstructures. The prompt generation demonstrates the LLM's capability to conceptualize new materials-inspired design principles, which have the potential to be synthesized into physical structures. Specifically, the innovative designs are guided by the decomposed principles generated through prompt generation, which are directly relevant to materials science. Besides, LLM-generated prompts, new design possibilities can be explored that may not have been intuitive through traditional design processes. This approach bridges the gap between conceptual design and practical material development, and shows how AI can assist in design informatics by accelerating concept exploration and identifying viable design features.

While slight variations of prompting are used to yield different examples (each result presented here includes a detailed presentation of all features), the general goal is to think about design principles that we can extract from combining different biological materials. For instance, we ask the model to think about ways to combine design elements from spider silk and collagen to make a strong, lightweight but tough material, and to also incorporate design cues from leaf microstructures. The approach can be used to focus directly on material microstructures but can also be used to yield cross-domain results, such as architectural ideas or city design. We have provided several design examples that span a range of applications, from structural to urban design, to illustrate the versatility and potential of

bio-inspired design principles. The examples showcase how models trained on biomimetic materials can conceptualize structural motifs commonly observed in hierarchical composites and biomaterial-inspired structures. Furthermore, by presenting visual representations of bio-inspired designs, we demonstrate the capability of the LLM to bridge conceptual design with potential real-world synthesis in the field of materials science.

In our first example, we prompt the `lamm-mit/SmolLM-Base-1.7B-CPT-SFT-DPO` model as follows (Fig. 19 shows the entire conversation):

Conversation template used with: `lamm-mit/SmolLM-Base-1.7B-CPT-SFT-DPO`

System: Your are a helpful and friendly assistant. You excel in combining disparate concepts and develop well thought-out designs of organic, bio-inspired material microstructures.

User: First, think about specific design principles you can extract from combining dragline silk and collagen fibers to make a strong, lightweight but tough material. Incorporate design cues from leaf microstructure.

Assistant: ...

User: Use these thoughts to develop an image generation prompt for an AI system like Stable Diffusion. Instead of using technical terms, you carefully describe the image using geometric features such as lines, shapes, circles, holes, porous features, patterns, and so on.

Assistant: ...

User: Give me one short prompt that I can use in my AI generator, but be specific about the features to be incorporated and how.

Assistant: Here's a concise prompt to get you started:

"Generate a futuristic, eco-friendly architectural concept utilizing a biomimetic composite material that integrates the structural efficiency of spider silk with the adaptive porosity of plant tissues. Utilize the following key features:

* Fibrous architecture inspired by spider silk, represented by sinuous lines and curved forms.
* Interconnected, spherical nodes reminiscent of plant cell walls, emphasizing growth and adaptation.
* Open cellular structures echoing the permeable nature of plant leaves, suggesting dynamic exchanges and self-regulation capabilities.
* Gradations of opacity and transparency inspired by the varying densities found in plant tissues, highlighting functional differentiation and multi-functionality."

This prompt encourages the AI generator to incorporate specific visual elements drawn from nature, ensuring a cohesive and engaging final output.

A collection of images produced in this way is shown in Fig. 20. The resulting images resemble a visionary architectural concept where biomorphic structures seamlessly blend with nature, creating a futuristic and sustainable environment. The designs are inspired by natural forms such as honeycombs, coral, and cellular structures, characterized by fluid, curving lines and intricate lattice-like frameworks that evoke the organic world. These structures are integrated with greenery, emphasizing harmony with the environment and suggesting the use of innovative, eco-friendly materials. The open, flowing spaces with large archways and natural light emphasize a connection with the outdoors, creating a sense of tranquility and well-being. As can be seen in some of the images, human figures within these spaces highlight the livability and community-centric design, suggesting a vision where technology and nature coexist harmoniously. The images present a forward-looking approach to architecture, where sustainability, esthetics, and advanced techniques converge to create a new paradigm for living and public spaces. A close inspection of the resulting designs further suggests a clear emergence of the leaf microstructure patterns, a result of the prompt and the fine-tuned ability of the generative model to incorporate this particular design idea.

These design ideas have the potential to be implemented, with specific functionality emphasized. Real-life examples of bio-inspired architectural designs include "The Hive", a building on the NUT campus designed by Heatherwick Studio. Its honeycomb structure mirrors the cellular organization of hives, providing a modified modular hexagonal form that improves structural efficiency, particularly by optimizing airflow and ventilation systems. Another compelling example is "Little Island" in New York, which is designed to resemble floating leaves and create a dynamic urban landscape. The design shows resilience to climate change, offering flexibility in adapting to changes in water levels.

As illustrated in these examples, bio-inspired designs in architecture go beyond mere esthetics by incorporating natural elements into urban facilities, often enhancing functionality. Specifically, as shown in Fig. 20, these designs have the potential to improve sustainability and energy efficiency by optimizing material usage and structural topology compared to conventional designs. For instance, the designs in row three reflect design simplicity and potential material saving, incorporating non-uniform cellular structures tailored for different space usages. The structural design also satisfies the load path analysis, with larger column sizes at lower floors compared to the top ones, which also shows the load-bearing capacity enhancement. Moreover, energy efficiency can be improved by optimizing heating and cooling systems, as seen in designs that mimic leaf veins, such as the first design in row one. The second design in row five demonstrates a bio-inspired roof design, with beam thickness inspired by leaf veins to maximize lighting and energy efficiency. Additionally, these designs enhance the user experience by integrating natural elements into urban facilities, such as bridges and pathways within the designs. However, to actualize these designs in real-life applications, further research is needed, which includes studying wind effects and façade analysis considering the integration of extended plants and ensuring the loading and structural integrity, as well as selecting materials that are compatible with greenery.

Furthermore, bio-inspired structures featuring cellular patterns have the potential to be further enhanced with topology optimization for various purposes, such as optimizing material usage while maintaining structural integrity and architectural features. We note that these bio-inspired designs reveal innovative design approaches for environmental integration and sustainability improvements, offering a promising start for exploring the vast design space, although more research is needed to fully validate the application in the real world.

In our second example, we prompt the `lamm-mit/mistral-7B-v0.3-Base-CPT-SFT-DPO` model as follows (Fig. 21 shows the entire conversation):

A collection of images produced in this way is shown in Fig. 22. The images showcase close-up views of a novel form of a biological material, highlighting their intricate vein patterns and microstructures. The leaves exhibit a variety of geometric vein arrangements, ranging from polygonal to radial patterns, all sharply contrasted against the green leaf surfaces by bright yellow to gold veins. The diversity in leaf shapes and vein structures offers a variety of structural options, including a prominent feature of spider-web-like motifs. Several of the leaf patterns in the image resemble spider webs or spider-web-like structures. The veins in these particular leaves form radial patterns that converge toward a central point, much like the structure of a 2D orb spider web. This similarity is especially pronounced in some of the leaves where the vein network is more symmetrical and evenly spaced, creating a web-like appearance. These spider-web-like patterns, some of which resemble projections of various 3D webs such as cobwebs and sheet webs, add an interesting visual for studies related to natural design and biomimicry. A close inspection of the resulting images shows that textures and depth are captured in fine detail. We find that the soft yet focused lighting accentuates these patterns.

Conversation template used with: `lamm-mit/mistral-7B-v0.3-Base-CPT-SFT-DPO`

System: Your are a helpful and friendly assistant. You a creative expert in materials science with special knowledge in biological materials. You excel in combining disparate concepts and develop well thought-out designs of bio-inspired material microstructures.

User: First, think about design principles you can extract from combining spider silk and collagen to make a strong, lightweight but tough material. Incorporate design cues from leaf microstructure.

Assistant: ...

User: Use these thoughts to develop an image generation prompt, incorporating the keyword 'leaf microstructure' when appropriate. Instead of using technical terms, you carefully describe the image using generic features such as lines, shapes, circles, holes, patterns, and so on.

Assistant: ...

User: Give me one short prompt that I can use in my AI generator.

Assistant: Specifically for a direct and concise image generation prompt related to the bio-inspired material microstructure, here's a one-liner:

"Generate an image of a golden spider web network intertwined with collagen veins, forming a dynamic, leaf-inspired microstructure amidst a lush green background."

Figure 23 shows a few additional sample images specifically prompting the model to develop urban design ideas based on a set of biological materials, including spider silk, collagen, and leaves, developed by `lamm-mit/SmolLM-Base-1.7B-CPT-SFT-DPO`. The images presented illustrate a conceptual approach to urban design that synthesizes advanced architectural techniques with principles of ecological integration and sustainability. The structures exhibit biomimetic design, characterized by their spiraling, organic forms that mimic natural patterns, possibly influenced by leaf microstructures. This design approach aligns with the principles of biophilic architecture[44], which aims to reconnect urban environments with the natural world by incorporating natural elements into the built environment.

Several similar architectural designs have been realized, validating the potential for these generated concepts. Examples include Azabudai Hills, a district in Tokyo featuring curving planted rooftops; CapitaSpring, a skyscraper in Singapore with orthogonal strips of plants embedded in its façade; and the Vertical Forest, designed by Stefano Boeri Architetti, which integrates residential buildings with diverse greenery. However, there remains significant potential to explore and expand these ideas within extended urban systems, for instance, how to interconnect individual structures, ensuring seamless integration with other facilities and alignment with an overall urban planning strategy.

The integration of nature into architecture has roots in movements such as organic architecture, as advocated by Frank Lloyd Wright[45], where the harmony between human habitation and the natural world is paramount. The designs proposed here, however, push this concept further by embedding extensive greenery directly into the architectural framework, creating vertical gardens and green terraces that are integral to the building's structure rather than ancillary elements. Moreover, the buildings are interconnected through elevated walkways, which not only facilitate human movement but also promote ecological connectivity, potentially serving as corridors for urban wildlife and contributing to biodiversity. This interconnectedness suggests a systems-thinking approach to urban design, where the built environment is considered part of a larger ecological network rather than an isolated entity.

The design concepts represented in these images could perhaps be seen as a potential paradigm shift in urban planning, where we move beyond sustainability to focus on regenerative design. This approach aims to create urban environments that not only minimize ecological impact but actively restore and enhance the natural environment. Such a model could potentially represent a significant advancement in urban ecology, proposing a future where cities operate as living systems, integrated with and supportive of their surrounding ecosystems. More work would be necessary to explore this, but this example illustrates a use case where the methods developed here can guide creative research and technology developments.

Discussion

This study addressed fundamental questions in the fine-tuning of large language models (LLMs) for domain-specific knowledge, exploring how different optimization strategies and datasets influence model performance, and assessed effects of model size and capabilities. Our investigation focused on a host of techniques applied consistently across models/architectures and parameter numbers. These included Continued Pre-Training (CPT), Supervised Fine-Tuning (SFT), Direct

Fig. 15 | Conversation generated by the model. `lamm-mit/Llama3.1-8b-Instruct-CPT-SFT-ORPO-SLERP-Var_G`.

Conversation with: `lamm-mit/Llama3.1-8b-Instruct-CPT-SFT-ORPO-SLERP-Var_G`

System: Your are a helpful and friendly assistant. You are an expert in materials science with special knowledge in biological materials. You provide concise but accurate responses. You are an expert in combining different concepts, often not obviously related, and provide accurate accounts of correct relationships or analogies.

User: Collagen and leaves, discuss their relationship.

Assistant: While collagen and leaves may seem unrelated at first glance, they share some interesting connections.

1. **Structural Similarities**: Both collagen and leaf structures exhibit hierarchical organization. Collagen is a protein that forms the primary structural component of connective tissues like skin, tendons, and bones. It has a helical structure that self-assembles into fibrils, which then form fibers and eventually tissues. Similarly, leaves have a hierarchical structure, starting from individual cells to complex vascular networks.

2. **Mechanical Properties**: Collagen provides strength and elasticity to tissues due to its unique triple-helix structure. Leaves also need to withstand mechanical stresses such as wind and rain. The arrangement of cellulose microfibrils within plant cell walls gives them tensile strength and flexibility, similar to how collagen's structure contributes to tissue resilience.

3. **Biological Functionality**: In living organisms, collagen plays a crucial role in maintaining tissue integrity and facilitating movement (e.g., in tendons). Leaves, on the other hand, perform photosynthesis, absorbing sunlight to produce energy for plants through chlorophyll. However, both components contribute to the overall health and function of the organism - collagen supports bodily functions, while leaves support life by producing oxygen and organic compounds.

4. **Materials Science Applications**: Understanding the properties of collagen can inform biomaterials design, particularly in medical applications where tissue repair or replacement is needed. Similarly, studying the mechanics of leaves can inspire the development of lightweight yet strong materials for engineering applications.

In summary, while collagen and leaves serve distinct biological purposes, they share commonalities in terms of structural organization, mechanical properties, and potential applications in materials science.

User: Thank you! Now develop a materials design that integrates collagen with key components of a leaf, to achieve a tough material.

Assistant: Designing a material that combines the beneficial properties of collagen with those of leaves requires understanding the unique characteristics of each component. Here's a hypothetical approach:

Material Design: "Leaf-Collagen Hybrid"

Key Components:

1. **Collagen Fibrils**: These would be the primary structural elements, providing strength and elasticity.
2. **Cellulose Microfibrils**: Inspired by the leaf's cell wall structure, these would add rigidity and resistance to deformation.
3. **Chloroplast-Inspired Nanoparticles**: Incorporating nanoparticles that mimic the light-absorbing capabilities of chloroplasts could enhance the material's ability to harness energy from external sources.

Manufacturing Process:

1. **Self-Assembly**: Collagen fibrils would be allowed to self-assemble into a network, mimicking natural tissue formation.
2. **Nanoparticle Integration**: Chloroplast-inspired nanoparticles would be dispersed throughout the collagen matrix, potentially using techniques like electrostatic deposition or chemical bonding.
3. **Cellulose Microfibril Reinforcement**: Cellulose microfibrils would be embedded within the collagen matrix, either through mechanical mixing or chemical cross-linking.

Properties:

1. **Toughness**: The combination of collagen's elasticity and cellulose's rigidity should result in a material that can absorb significant amounts of energy without breaking.
2. **Lightweight**: The use of cellulose microfibrils and nanoparticles could help reduce the material's weight while maintaining strength.
3. **Energy Harvesting**: The chloroplast-inspired nanoparticles might enable the material to convert environmental energy (e.g., light) into usable power, making it suitable for applications like wearable technology or biomedical devices.

Potential Applications:

1. **Biomedical Implants**: The Leaf-Collagen Hybrid could be used in implants that require both strength and biocompatibility, such as artificial joints or tissue scaffolds.
2. **Wearable Technology**: Its energy harvesting capabilities make it an attractive candidate for smart clothing or accessories that can power small electronics.
3. **Environmental Monitoring**: The material's ability to respond to environmental stimuli could be leveraged for monitoring air quality or detecting changes in temperature.

This hypothetical design illustrates how integrating key components of collagen and leaves could lead to the creation of a tough, lightweight, and energy-harvesting material with diverse applications.

User: Summarize these design features in JSON format.

Assistant: ```json

```
{
    "name": "Leaf-Collagen Hybrid",
    "components": [
        {
            "type": "collagen fibrils",
            "function": "primary structural element, provides strength and elasticity"
        },
        {
            "type": "cellulose microfibrils",
            "function": "rigidity and resistance to deformation"
        },
        {
            "type": "chloroplast-inspired nanoparticles",
            "function": "energy harvesting capability"
        }
    ],
    "manufacturing_process": [
        {
            "step": "self-assembly of collagen fibrils",
            "description": "mimics natural tissue formation"
        },
        {
            "step": "integration of nanoparticles",
            "description": "electrostatic deposition or chemical bonding"
        },
        {
            "step": "reinforcement with cellulose microfibrils",
            "description": "mechanical mixing or chemical cross-linking"
        }
    ],
    "properties": [
        {
            "property": "toughness",
            "value": "high energy absorption capacity"
        },
        {
            "property": "lightweight",
            "value": "reduced weight while maintaining strength"
        },
        {
            "property": "energy harvesting",
            "value": "conversion of environmental energy into usable power"
        }
    ],
    "applications": [
        {
            "application": "biomedical implants",
            "description": "artificial joints or tissue scaffolds requiring strength and biocompatibility"
        },
        {
            "application": "wearable technology",
            "description": "smart clothing or accessories powering small electronics"
        },
        {
            "application": "environmental monitoring",
            "description": "detecting changes in temperature or air quality"
        }
    ]
}
```

This JSON representation summarizes the design features of the Leaf-Collagen Hybrid, including its components, manufacturing process, properties, and potential applications.

Preference Optimization (DPO), and Odds Ratio Preference Optimization (ORPO). The goal was to determine how these methods impact the specialization of LLMs, particularly in the context of engineering or science domains.

A key finding of our research is that model scale plays a crucial role in the efficacy of fine-tuning strategies. Larger models, such as those with 7B and 8B parameters, not only exhibited substantial improvements in domain-specific tasks but also showed the emergence of novel capabilities-an outcome not observed in smaller models like the 1.7B parameter SmolLM model. This observation, as shown in a comparative plot (Fig. 13) suggests a threshold effect, where the benefits of advanced optimization techniques, including model merging through SLERP (Spherical Linear

Conversation with: `lamm-mit/mistral-7B-v0.3-Base-CPT-SFT-DPO`

System: Your are a helpful and friendly assistant. You are an expert in materials science with special knowledge in biological materials. You provide concise but accurate responses. You are an expert in combining different concepts, often not obviously related, and provide accurate accounts of correct relationships or analogies.

User: Collagen and leaves, discuss their relationship.

Assistant: As an artificial intelligence, I'm glad to assist you in exploring the relationship between collagen and leaves, specifically by highlighting the following aspects:

Collagen: Collagen is a fibrous structural protein that is the most abundant protein in mammals. It is a key component of various tissues such as tendons, ligaments, skin, bones, cartilage, and blood vessels. Collagen molecules consist of three polypeptide chains (alpha-1, alpha-2, and alpha-3) arranged in a triple helix structure. These chains are composed of repeating sequences of amino acids, including glycine, proline, and hydroxyproline.

Leaves: Leaves are specialized organs in plants that play a crucial role in photosynthesis, respiration, and transpiration. They have several layers of cells, each with specific functions. The outermost layer, known as the cuticle, is a waxy layer that helps reduce water loss through transpiration. Beneath the cuticle lies the epidermis, which contains the stomata (tiny pores).

The connection between collagen and leaves may seem unrelated at first glance, but let me explain how some analogies can be drawn:

1. **Structural similarities**: Both collagen and leaves exhibit hierarchical structures. Collagen has a triple helical arrangement of polypeptide chains, while leaves have a layered structure consisting of the cuticle, epidermis, mesophyll, and vascular tissue. Both structures contribute to the mechanical strength and resilience of the respective organisms.

2. **Mechanical properties**: Collagen exhibits high tensile strength and elasticity due to its unique triple helix structure. Similarly, leaves possess remarkable mechanical properties due to their cellular arrangement and composition. For example, the sclerenchyma cells in the leaf veins provide support and rigidity to the leaf structure.

3. **Adaptability**: Collagen adapts to changes in the body's physiological conditions, such as during wound healing or tissue regeneration. In a similar manner, leaves adapt to environmental factors like wind, rain, and temperature fluctuations. This adaptation is achieved through changes in the cellular structure and arrangement.

4. **Functionality**: Collagen plays a vital role in maintaining the integrity and function of various tissues in the body. Leaves perform essential functions for plant growth and survival, such as gas exchange, nutrient absorption, and light capture.

In conclusion, while collagen and leaves may appear dissimilar on the surface, they share many underlying characteristics that make them both fascinating examples of biological materials. By understanding these similarities, we can gain insights into the fundamental principles that govern the design and function of natural systems.

User: Thank you! Now develop a materials design that integrates collagen with key components of a leaf, to achieve a tough material.

Assistant: To create a tough material integrating collagen with key components of a leaf, I will propose a bioinspired composite material concept that combines specific structural and mechanical features from both collagen and leaf tissues.

Material Components:
1. **Collagen Fibrils**: We will utilize collagen fibrils as the primary reinforcing fibers in our composite material. Collagen fibrils, known for their exceptional tensile strength and elasticity, will serve as the load-bearing backbone of the composite.

2. **Cellulose Nanofibers (CNFs)**: CNFs, derived from the cell walls of plants, including leaves, offer excellent mechanical properties like stiffness, strength, and toughness. Their nanoscale dimensions allow them to form strong interfacial bonding with other materials, enhancing the overall performance of the composite.

3. **Silk Fibroin**: Silk fibroin, another biopolymer, provides additional mechanical strength and flexibility. Its molecular structure allows it to form robust interactions with collagen and CNFs, contributing to the overall cohesion and toughness of the composite.

4. **Nanocellulose-Based Matrix**: A matrix made from nanocellulose, such as bacterial cellulose, will be used to bind the reinforcing fibers together. This matrix offers superior mechanical properties compared to traditional polymer matrices, allowing the composite to maintain its toughness under stress.

Design Strategy:
1. **Preparation of Collagen Fibrils**: Collagen fibrils will be extracted and purified using established methods.

2. **Production of Cellulose Nanofibers (CNFs)**: CNFs will be isolated from leaf sources, such as hemp or flax, using techniques like acid hydrolysis or mechanical grinding.

3. **Integration of Silk Fibroin**: Silk fibroin will be dissolved and blended with the collagen and CNF solutions to enhance the composite's overall mechanical properties.

4. **Formulation of Nanocellulose-Based Matrix**: A nanocellulose-based matrix will be prepared by dissolving bacterial cellulose in suitable solvents.

5. **Mixing and Casting**: All the components will be mixed thoroughly to ensure homogeneity before casting into molds.

6. **Drying and Post-Processing**: After casting, the composite material will undergo drying and post-processing steps, such as crosslinking, to improve its mechanical stability and durability.

This proposed composite material integrates the unique mechanical properties of collagen, CNFs, silk fibroin, and nanocellulose-based matrices. The combination of these components results in a tough, durable, and potentially biodegradable material that could find applications in various fields, including biomedical engineering, protective gear, and advanced composites.

User: Summarize these design features in JSON format.

Assistant: Under specific requests for JSON format summarization:

```json
{
    "title": "Bioinspired Tough Composite Material Integrating Collagen, Cellulose Nanofibers, and Silk Fibroin",
    "insights": [
        "Combining collagen, cellulose nanofibers (CNFs), and silk fibroin creates a composite material with enhanced mechanical properties.",
        "A nanocellulose-based matrix improves the overall cohesion and toughness of the composite.",
        "The proposed composite material demonstrates potential applications in various fields, including biomedical engineering and protective gear."
    ],
    "facts": [
        "Collagen fibrils serve as the primary reinforcing fibers.",
        "Cellulose nanofibers (CNFs) offer excellent mechanical properties and facilitate strong interfacial bonding.",
        "Silk fibroin contributes additional mechanical strength and flexibility.",
        "A nanocellulose-based matrix binds the reinforcing fibers together.",
        "The composite material can be produced via mixing, casting, and drying processes."
    ],
    "details": {
        "Title": "Bioinspired Tough Composite Material Integrating Collagen, Cellulose Nanofibers, and Silk Fibroin",
        "Insights": [
            "Collagen + CNFs + SF + Nanocellulose-based matrix = Enhanced mechanical properties",
            "The composite material exhibits potential applications in various fields due to its unique mechanical properties."
        ],
        "Facts": [
            "Collagen fibrils are extracted and purified.",
            "Cellulose nanofibers (CNFs) are isolated from leaf sources.",
            "Silk fibroin is dissolved and blended with collagen and CNF solutions.",
            "A nanocellulose-based matrix is prepared by dissolving bacterial cellulose.",
            "The composite material is mixed, cast, dried, and post-processed to improve mechanical stability and durability."
        ]
    }
}
```

Fig. 16 | Conversation generated by the model. `lamm-mit/mistral-7B-v0.3-Base-CPT-SFT-DPO`.

Interpolation), become significantly more pronounced as model size increases. There are notable differences between the Llama and Mistral family of models that deserve further investigation. However, such investigations are hampered by the lack of detailed insights into the datasets used for training (during pre-training and fine-tuning) as well as a lack of details on specific training approaches. A summary of the key insights, including the most effective approach, is shown in Table 4.

The comparison between the `Llama-3.1-8b` and `Mistral-7B-v0.3` models highlighted how SLERP, when combined with SFT, DPO and ORPO, can effectively unlock synergistic properties, leading to refined

Conversation with: `lamm-mit/mistral-7B-v0.3-Base-CPT-SFT-SLERP`

System: Your are a helpful and friendly assistant. You are an expert in materials science with special knowledge in biological materials. You provide concise but accurate responses. You are an expert in combining different concepts, often not obviously related, and provide accurate accounts of correct relationships or analogies.

User: Collagen and leaves, discuss their relationship.

Assistant: Collagen is a protein that is found in the connective tissues of animals, including humans. It provides strength and structure to these tissues. Leaves, on the other hand, are the photosynthetic organs of plants. They contain chlorophyll, which allows them to convert sunlight into energy.

The relationship between collagen and leaves can be seen in the way they both function as structural components in their respective organisms. Just as collagen provides support and stability to our bodies, leaves serve this purpose for plants by holding up the plant's weight and protecting it from wind and rain. Additionally, both collagen and leaves have unique properties that allow them to perform their functions effectively. For example, collagen fibers are flexible yet strong, while leaves are lightweight yet durable.

In summary, while collagen and leaves may seem very different at first glance, they share some similarities in terms of their roles as structural components in their respective organisms. Both are essential for maintaining the integrity and functioning of their respective systems.

User: Thank you! Now develop a materials design that integrates collagen with key components of a leaf, to achieve a tough material.

Assistant: To create a tough material using collagen and key components of a leaf, we could consider designing a composite material that incorporates the following elements:

1. Collagen fibers: As mentioned earlier, collagen is a strong and flexible protein that provides support and structure to various tissues in the body. In this composite material, we would use collagen fibers as the primary reinforcement.

2. Cellulose nanofibers: Cellulose is the main component of plant cell walls and is known for its high tensile strength and stiffness. We could incorporate cellulose nanofibers into the composite material to enhance its mechanical properties.

3. Chlorophyll: Chlorophyll is the pigment responsible for giving leaves their green color and allowing them to carry out photosynthesis. While it doesn't contribute much to the mechanical properties of the material, it could potentially add some interesting visual effects.

4. Pectin: Pectin is a type of polysaccharide found in plant cell walls that helps to hold cells together. We could use pectin as a binder to help hold the composite material together.

To create the composite material, we would mix the collagen fibers, cellulose nanofibers, chlorophyll, and pectin together in a solution. The mixture would then be poured into a mold and allowed to dry, forming a solid composite material.

This composite material would likely exhibit high tensile strength and flexibility due to the presence of collagen fibers and cellulose nanofibers. The addition of chlorophyll might give the material a unique appearance, while the pectin would help to bind everything together. Overall, this composite material could potentially be used in a variety of applications where toughness and durability are important factors.

User: Summarize these design features in JSON format.

Assistant:
```
{
    "material_components": [
        "collagen_fibers",
        "cellulose_nanofibers",
        "chlorophyll",
        "pectin"
    ],
    "binder": "pectin",
    "reinforcement": "collagen_fibers",
    "enhancements": [
        "cellulose_nanofibers"
    ]
}
```

Fig. 17 | Conversation generated by the model. `lamm-mit/mistral-7B-v0.3-Base-CPT-SFT-SLERP`.

performance outcomes that cluster distinctly in our analysis. However, these improvements were not mirrored in the smaller `SmolLM-1.7B` models, which showed a deterioration of performance under model merging, underscoring the importance of scale in realizing the full potential of fine-tuning techniques. Figure 24 shows an overview of the significant performance gains achieved through fine-tuning across different model families, emphasizing how larger models like Llama and Mistral benefit more substantially from these advanced techniques in terms of the overall scale of performance reached.

As shown in Fig. 24A, fine-tuning consistently enhances model performance, with Llama leading in absolute performance scores. However, Fig. 24B reveals an interesting trend: Despite SmolLM's lower baseline which is likely due to the inherent limitations of its smaller parameters count and reduced access to high-dimensional parameter spaces compared to larger models, its relative improvement from fine-tuning is the most pronounced, illustrating how smaller models, while not reaching the peak performance of their larger counterparts, can still gain substantially from targeted optimization strategies. This differential improvement pattern underscores the critical role of model size in both the efficacy and scope of fine-tuning processes. Consequently, while the larger Llama and Mistral models are better equipped to leverage the full spectrum of fine-tuning techniques like DPO and ORPO combined with SLERP, SmolLM demonstrates that even smaller-scale models can achieve meaningful performance boosts, albeit with limitations imposed by their smaller capacity.

Figure 25 shows performance over the number of pre-training tokens (note, 1 trillion for SmolLM, 8 trillion for Mistral, and 15 trillion were used for Llama-3.1). While this is an approximate analysis as each model/architecture has unique features, training strategies, and datasets associated with it, but an overall trend can be observed that more pre-training tends to yield better performance, but also diminishes relative improvements that can be achieved. This agrees with the general consensus that for transformer-based models, there seem to be diminishing returns as training data and model size are increased, implying that as performance gets better, exponentially larger and better datasets and larger models must be used.

The results reported here contribute to the ongoing discourse on the optimal strategies for fine-tuning LLMs, particularly in specialized domains where precise and sophisticated understanding is required. The results suggest that while CPT and SFT are foundational, the application of advanced optimization methods like DPO and ORPO, especially when coupled with model merging techniques like SLERP, are critical for achieving high-performance outcomes in larger models. The data shows that there is room to optimize performance for each model, and some basic understanding needs to be developed of how different base models perform and how they can best be improved. Overall, the use of CPT, SFT, and ORPO/DPO followed by model merging can be seen as a viable approach for strong domain performance. The use of CPT and the other steps tends to yield much better results than simpler approaches such as LoRA, albeit, at a much higher computational cost. For smaller models, model merging may not be advisable, albeit we would recommend experimentation for model architectures to explore this particular feature and also identify the threshold of model size in which emergent capabilities prominently appear.

Conversation with: lamm-mit/SmolLM-Base-1.7B-CPT-SFT-DPO

System: Your are a helpful and friendly assistant. You are an expert in materials science with special knowledge in biological materials. You provide concise but accurate responses. You are an expert in combining different concepts, often not obviously related, and provide accurate accounts of correct relationships or analogies.

User: Collagen and leaves, discuss their relationship.

Assistant: An intriguing pair! Collagen and leaves may seem like an unusual combination at first glance, but they do share some fascinating connections.

1. **Structural Support**: Both collagen and leaves have evolved to serve as structural support systems for their respective organisms. In the case of leaves, they provide mechanical stability against environmental forces such as wind, rain, and gravity. Similarly, collagen provides strength and integrity to various tissues within animals by forming a framework that supports other cellular components.

2. **Cellular Architecture**: The hierarchical structure of both collagen and leaves is similar - from macroscopic (leaves) to microscopic (collagen fibers). This architecture allows them to efficiently distribute stress and maintain functionality across multiple scales. For example, leaf veins branch out into smaller capillaries, which then feed into larger vessels, much like how collagen fibrils organize into lamellae, sheets, and ultimately form fiber bundles.

3. **Biomimicry**: The study of these natural structures has inspired innovative biomaterials and bioinspired designs. Researchers draw inspiration from the efficient load-bearing capabilities found in both leaves and collagenous tissues. For instance, synthetic composites mimicking the layered structure of leaves can be used in applications requiring high tensile strength and toughness, while engineered collagen scaffolds aim to replicate the complex organization seen in native tissues.

In summary, despite being seemingly unrelated, the connection between collagen's role in providing structural support and the functional design of leaves highlights the universal principles of material optimization observed throughout nature.

User: Thank you! Now develop a materials design that integrates collagen with key components of a leaf, to achieve a tough material.

Assistant: What a great question! Here's a novel materials design concept that combines the strengths of collagen with key elements of a leaf:

Material Name: Leafy-Coraline (LC) Composite

Principle: Inspired by the hierarchical structure of leaves and the exceptional mechanical properties of coral reefs, this composite material leverages the unique characteristics of collagen and cellulose nanocrystals (CNCs) to create a tough, lightweight, and adaptive material.

Composition:

1. **Collagen Fibrils (CF):** Assemble into microfibers using a wet spinning technique, incorporating a crosslinking agent to enhance mechanical stability. CF diameters range from 50-400 nm, ensuring flexibility and resilience.
2. **Nanocrystalline Cellulose (NCC):** Incorporated into the CF matrix through a solvent casting method, resulting in CNCs with dimensions around 8-16 nm. These crystalline structures improve interfacial bonding and overall mechanical performance.
3. **Alginate Adhesive:** Used as a binder to ensure proper dispersion of CF and NCC within the LC matrix. Alginate forms strong hydrogen bonds with both CF and CNC, enhancing adhesion and cohesion.

Design Features:

1. **Hierarchical Organization:** The CF/NCC/alginate matrix mimics the multiscale structure of leaves, featuring a fibrous network embedded in a more amorphous matrix. This arrangement optimizes mechanical properties, allowing for enhanced toughness and resistance to deformation.
2. **Adaptability:** The incorporation of CNCs enables the LC composite to exhibit shape memory effects, enabling it to recover its original configuration after deformation. This property makes the material suitable for dynamic applications where adaptability is crucial.
3. **Self-Healing:** The alginate adhesive facilitates self-healing behavior upon damage, promoting rapid recovery of mechanical integrity without external intervention.

Properties and Performance:

* Toughness modulus: ~70 MPa
* Elastic modulus: ~9 GPa
* Strength: ~100 MPa
* Strain hardening capacity: >20%
* Shape memory response: reversible recovery up to 90% strain

The Leafy-Coraline composite demonstrates remarkable mechanical robustness, coupled with the ability to adapt to changing conditions and self-heal when damaged. Such features make it an ideal candidate for advanced applications in fields like aerospace, biomedical devices, and sustainable infrastructure development.

User: Summarize these design features in JSON format.

Assistant: Here's a summarized JSON representation of the Leafy-Coraline composite design features:

```json
{
  "composition": {
    "microfibers": {
      "diameter": [50, 400],
      "material": "collagen",
      "adhesive": "alginate"
    },
    "nanocrystals": {
      "dimensions": [8, 16, 16],
      "material": "cellulose nanocrystals"
    }
  },
  "structure": {
    "hierarchical organization": true,
    "fiber network": true,
    "matrix composition": true
  },
  "properties": {
    "toughness_modulus": 70,
    "elastic_modulus": 9,
    "strength": 100,
    "strain_hardening_capacity": 20,
    "shape_memory_response": {
      "recovery_percentage": 90,
      "recovery_rate": 0.90
    }
  },
  "self-healing": true
}
```

This JSON represents the design features of the Leafy-Coraline composite, including its hierarchical structure, fiber network, matrix composition, and impressive mechanical properties along with the ability to self-heal under certain conditions.

Fig. 18 | Conversation generated by the model. lamm-mit/SmolLM-Base-1.7B-CPT-SFT-DPO.

There are numerous other parameters and angles to be explored. First, the effect of prompting. We kept the prompting consistent across all experiments. However, each model or model family will likely respond differently to prompts, and it is possible that additional performance gains can be achieved via prompt engineering. We deliberately did not address this issue here as we wanted to focus on an overall, consistent comparison. In some preliminary assessments, we did notice that for the Mistral models, the CPT-SFT-DPO series resulted in rather low benchmark performance, where a closer inspection suggested that this may be due to limited ability to follow directions, and lengthy answers. Our benchmarks are consistent in

Table 2 | Summary of model performance, strengths, weaknesses, and design strategies based on conversations related to collagen and leaves

Model	Key strengths	Weaknesses	Observations
lamm-mit/Llama3.1-8b-Instruct-CPT-SFT-DPO	Detailed and comprehensive design; strong scientific grounding	May be overly complex for some applications; focuses heavily on structural components, potentially missing opportunities for simpler solutions	The model excels in providing a thorough and detailed analysis. The robust material design effectively integrates various components, showcasing strong scientific reasoning. However, the complexity of the design may limit its practicality in certain applications. The JSON summary is precise, but the depth of detail could be overwhelming in less technical contexts.
lamm-mit/Llama3.1-8b-Instruct-CPT-SFT-ORPO-SLERP-Var_G	Concise and inventive; broader application vision	Lacks depth in certain areas; the brevity may sacrifice some detail in the explanation of the material properties	This model offers a concise yet insightful analysis. Its focus on innovation is evident in the integration of chloroplast-inspired nanoparticles for energy harvesting, which is unique among the models. However, the lack of depth in the explanation of mechanical properties could be a drawback in more technical discussions. The JSON representation is clear and well-structured, reflecting the model's focus on innovative applications.
lamm-mit/mistral-7B-v0.3-Base-CPT-SFT-DPO	Comprehensive integration of components; strong mechanical focus	Potentially too focused on mechanical properties, which might limit creativity in design	The model provides a detailed and comprehensive approach, particularly in its focus on the mechanical properties of the proposed material. The integration of multiple materials, such as cellulose nanofibers and silk fibroin, is handled well, though the emphasis on mechanical properties may limit the exploration of other innovative aspects. The JSON summary effectively captures the design's complexity, though it may be seen as overly technical for broader audiences.
lamm-mit/mistral-7B-v0.3-Base-CPT-SFT-SLERP	Simple and straightforward; clear and accessible design	Lacks depth and innovation; may be seen as too simplistic compared to other models	This model provides a straightforward and accessible design, focusing on essential components like collagen fibers and cellulose nanofibers. While the simplicity makes it accessible, it also limits the depth and creativity of the design. The JSON summary is basic but clear, making it suitable for less technical audiences but potentially insufficient for more advanced applications.
lamm-mit/SmolLM-Base-1.7B-CPT-SFT-DPO	Highly creative and detailed; innovative features like self-healing	Complexity might make it challenging to implement; could be seen as too speculative for practical use	The model stands out for its creativity and innovation, particularly in its inclusion of self-healing and shape-memory features. These advanced characteristics make it suitable for high-end applications, though the complexity of the design may pose challenges in practical implementation. The JSON summary is comprehensive, effectively conveying the unique aspects of the design, but the speculative nature of some features may limit its immediate applicability.

Evaluation conducted using GPT-4o.

that they test not only domain knowledge but also how closely models follow directives to answer in a certain way (single word, as done here). Further experiments with the Mistral CPT-SFT-DPO series could entail using a more diverse dataset during DPO, including additional instruction-following foci. These and other variations are left to future work.

We presented various materials-specific applications of the models that go beyond question-answer benchmarking, focusing on general reasoning capabilities to integrate complex, disparate biological materials concepts, step-by-step analysis, structured outputs, and others. Results of consistently prompted conversations with various models were shown in Figs. 14–18. These conversations of a human user with models provided important insights into the different strengths and weaknesses of the models in realistic use-cases.

Building on this, focusing on another real-world use case of our models, we prompted the LLMs to develop image generation prompts, showing a powerful use case of the LLMs in conjunction with a pipeline of models interacting via agentic modalities (see, e.g., Figs. 20, 22, and 23 ranging from futuristic architecture, bio-inspired material microstructures, to urban/cityscape design). The design examples underscore the transformative potential of bio-inspired design principles, demonstrating how models trained on biomimetic materials can inform the conceptualization of structural motifs characteristic of hierarchical composites and biomaterial-inspired systems. These insights pave the way for integrating advanced LLM capabilities into materials science, bridging innovative design with practical synthesis. In addition, these studies could be further expanded by incorporating a feedback loop, where in a multi-agent framework, images generated can be analyzed by vision-LLMs (e.g., Cephalo[46]) and the insights fed back into the LLMs for improved reasoning and adjustments.

Future research could further go deeper into the mechanisms behind the observed emergent behaviors, exploring how they can be harnessed across different architectures and domains, perhaps using tools from statistical mechanics or thermodynamics that may offer a fundamental perspective of how multi-particle systems behave and how these behaviors can be modeled. Other work to focus on interpretable insights could help also, following recent work[47]. Additionally, understanding the limitations and potential of smaller models remains an important area of inquiry, particularly as we seek to fine-tune LLMs for specific tasks without the extensive computational resources required for larger models. Other avenues include scale-up of the experiments to 70B or 405B models in the Llama series. Given the lessons learned from the present paper, a good strategy could be to do CPT on the Instruct model and then merge with the original Instruct model using SLERP. On the other side, more research could be done with the small LLMs. Our result based on SmolLM is interesting as they provide a model with reasonable performance on a tiny scale; this model could, for instance, become an effective tool when combined with in-context learning such as retrieval-augmented generation (RAG)[48].

Table 3 | Summary of model performance with individual criterion scores, total intelligence score, average score, and normalized intelligence score

Model	Depth of reasoning	Creativity	Clarity	Quantitative predictions	Total score (Max 40)	Average score (Max 10)	Normalized intelligence score
lamm-mit/SmolLM-Base-1.7B-CPT-SFT-DPO	10	10	9	9	38	9.5	10.0
lamm-mit/Llama3.1-8b-Instruct-CPT-SFT-DPO	9	8	9	7	33	8.25	8.7
lamm-mit/Llama3.1-8b-Instruct-CPT-SFT-ORPO-SLERP-Var_G	8	9	8	6	31	7.75	8.2
lamm-mit/mistral-7B-v0.3-Base-CPT-SFT-DPO	8	7	8	7	30	7.5	7.9
lamm-mit/mistral-7B-v0.3-Base-CPT-SFT-SLERP	6	6	7	5	24	6.0	6.3

We assign scores from 1 to 10 for each criterion, including (1) Depth of reasoning: how well the model explains concepts and connects ideas, (2) Creativity: the uniqueness and innovation in the material design, (3) Clarity: how clearly the model communicates its ideas, (4) Quantitative predictions: whether the model includes numerical or quantitative aspects in its response. Evaluation conducted using `GPT-4o`.

This study sheds light on the nuanced role of model scale and fine-tuning strategies in the development of domain-specific LLMs. By advancing our understanding of these dynamics, we move closer to unlocking the full capabilities of AI in specialized fields. The high degree of complexity across parameters and variables leaves this to be a challenging field of study that offers many opportunities for future research. More work could also be done in improving the datasets. We find that using a larger dataset is not necessarily beneficial for downstream performance. We also found similar results for the Mistral model, where we did not see a significant decrease but an almost identical performance. Combined with other recent studies, this indicates that data quality is a major issue that can be addressed by further distillation, processing and perhaps filtering data components for relevance. The use of DPO or ORPO is particularly intriguing as it offers avenues for improving scientific accuracy and aligning the model with particular styles of responses (e.g., reasoning in a systematic way, step-by-step). Dataset curation could target several of these aspects and more experiments could shed light on the effects of these on performance. In the same vein, the incorporation of visual cues (e.g., figures, plots, microstructures, etc.) as done in recent work[46] can be another source for data.

Materials and methods

We provide details about the materials and methods used to conduct this study.

Dataset curation and processing

The dataset used for training includes scientific papers from broad domains of biological materials, mechanics/mechanical properties, and spider silk. Earlier work focused on around 1000 papers in training; here, the dataset consists of the original training set and an additional set of ~4300 papers in the realm of spider silk. All of the training was done with this integrated dataset unless mentioned otherwise. A few experiments were done where training also included an extended dataset of ~3800 more papers on biological materials, see details below.

The dataset used for training includes scientific papers from broad domains of biological materials, mechanics/mechanical properties, and spider silk. Earlier work focused on around 1000 papers in training; here, the dataset consists of the original training set and an additional set of ~4300 papers in the realm of spider silk. All of the training was done with this integrated dataset unless mentioned otherwise. A few experiments were done where training also included an extended dataset of ~3800 more papers on biological materials, see details in the following sections.

Additionally, the quantity and quality of the dataset are crucial factors in the model training process. The curation and processing procedures outlined below ensure that the datasets are domain-specific, high-quality, and relevant. Specifically, for CPT, data quality is especially critical. As shown in Fig. 9, using an extended dataset with varied formats and defective text led to a decline in model performance. To overcome these challenges and ensure the accuracy, consistency, and efficiency of specialized LLMs, we implemented several key strategies: (1) Datasets were cleaned, structured, and formatted before training. (2) Model performance was rigorously evaluated using a tailored benchmark for the spider silk and bio-inspired materials domains, covering diverse question types and knowledge areas to ensure comprehensive validation. (3) A systematic training pipeline (as detailed in the "Results" section) was followed to maintain consistency at each stage. (4) Model merging was applied to integrate strengths from various training paths, enabling generalization and the emergence of new capabilities.

Firstly, we introduce a new dataset, the spider silk materials dataset, tailored to studies on spider silk. As the first step, we curated a set of PDF files using the method outlined in Fig. 26. The process involved four key steps:

1. Collecting 4520 papers from Web of Science,
2. Performing digital object identifier (DOI) lookups for missing entries using the `habanero.Crossref` Python client,
3. Screening for duplicates and irrelevant entries, and

Conversation with: `lamm-mit/SmolLM-Base-1.7B-CPT-SFT-DPO`

System: Your are a helpful and friendly assistant. You excel in combining disparate concepts and develop well thought-out designs of organic, bio-inspired material microstructures.

User: First, think about specific design principles you can extract from combining dragline silk and collagen fibers to make a strong, lightweight but tough material. Incorporate design cues from leaf microstructure.

Assistant: To create a novel biomimetic composite with exceptional mechanical properties, I'll draw inspiration from the unique structural features found in natural materials like spider dragline silk and leaf microstructure. By combining these two sources, I aim to replicate the desirable characteristics of both: strength, toughness, and light weight.

1. **Dragline Silk's Hierarchical Structure**: Dragline silk exhibits an intricate hierarchical structure composed of nanofibrils that form larger fibril bundles. This arrangement allows for efficient load distribution during deformation, contributing to its remarkable tensile strength. Incorporating this hierarchical organization into my design will help achieve similar performance while maintaining flexibility.

2. **Leaf Microstructure's Gradient Properties**: Leaf microstructure is characterized by gradual transitions between different tissue types (e.g., epidermis, mesophyll) which provide enhanced mechanical stability through stress redistribution. Integrating such gradient properties within my composite could lead to improved resistance against crack propagation and increased overall toughness.

3. **Collagen Fiber's Toughness**: Collagen fibers possess high tensile strength due to their triple helix configuration, allowing them to withstand significant forces without breaking. By incorporating collagen fibers into my design, I can enhance the composite's ability to absorb energy before failure occurs.

4. **Interpenetrating Network Formation**: To mimic the synergistic effects observed in nature where multiple components interact to produce superior properties, I propose forming an interpenetrating network (IPN) using both dragline silk and collagen fibers. This IPN would allow for effective stress transfer across different phases, leading to optimized mechanical behavior akin to that seen in natural composites.

By integrating these design principles inspired by dragline silk and leaf microstructure, along with the toughening mechanisms inherent in collagen fibers, I believe it's possible to craft a new class of biomimetic materials that exhibit outstanding mechanical properties comparable to those found in natural biological systems.

User: Use these thoughts to develop an image generation prompt for an AI system like Stable Diffusion. Instead of using technical terms, you carefully describe the image using geometric features such as lines, shapes, circles, holes, porous features, patterns, and so on.

Assistant: Here's a potential image generation prompt based on your request:

Prompt: "Generate a surreal, abstract composition featuring a biomimetic composite material that combines the structural integrity of spider silk and the adaptive porosity of plant tissues."

Description:

* The generated image should showcase a complex interplay of geometric features reminiscent of natural structures, including:
    + Lines and curves resembling the fibrous architecture of spider silk, evoking a sense of strength and resilience.
    + Circles and spheres representing the interconnected networks found in plant cells, symbolizing growth and adaptability.
    + Holes or voids mimicking the porous channels present in plant tissues, conveying a sense of permeability and responsiveness to environmental stimuli.
    + Porous regions suggestive of the open cellular structures typical of plant leaves, hinting at dynamic interactions and exchange processes.
    + Patterns and textures derived from the intersection of these geometric elements, creating a visually intriguing and dynamic composition.

The resulting image should capture the essence of biomimicry, blending the robustness of spider silk with the versatility of plant tissues, ultimately yielding a captivating representation of innovative material science.

User: Give me one short prompt that I can use in my AI generator, but be specific about the features to be incorporated and how.

Assistant: Here's a concise prompt to get you started:

"Generate a futuristic, eco-friendly architectural concept utilizing a biomimetic composite material that integrates the structural efficiency of spider silk with the adaptive porosity of plant tissues. Utilize the following key features:

* Fibrous architecture inspired by spider silk, represented by sinuous lines and curved forms.
* Interconnected, spherical nodes reminiscent of plant cell walls, emphasizing growth and adaptation.
* Open cellular structures echoing the permeable nature of plant leaves, suggesting dynamic exchanges and self-regulation capabilities.
* Gradations of opacity and transparency inspired by the varying densities found in plant tissues, highlighting functional differentiation and multi-functionality."

This prompt encourages the AI generator to incorporate specific visual elements drawn from nature, ensuring a cohesive and engaging final output.

Fig. 19 | Conversation generated by the model. `lamm-mit/SmolLM-Base-1.7B-CPT-SFT-DPO`, aimed to develop an image generation prompt. The interactions include several stages, such as ideation and principle identification, the creation of the image generation prompt, and the solicitation of a short version of the prompt.

4. Downloading 4323 papers through publisher application programming interfaces (APIs), manual downloads, and library requests.

A total of 4323 spider silk-related papers were downloaded in PDF format and used for training. 4520 papers were initially collected from the Web of Science search engine on April 17, 2024, using the keywords "spider silk". The search was limited to English-language articles published between 1900-01-01 and 2024-04-17. For the collected papers that were missing Digital Object Identifiers (DOIs), we used the `habanero.Crossref` Python client to interact with the CrossRef API, conducting DOI lookups based on the article titles, publication years, authors, and journals, where available. After initial collection, the paper entries were screened before downloading. Ultimately, 4323 papers were successfully downloaded in PDF format, resulting in a 95.6% yield. The remaining papers were not downloaded due to duplication, irrelevant content, or unavailability. Among the 4323 papers, 1603 were downloaded using the APIs of three publishers (420 from Wiley, 450 from Springer Nature, and 733 from Elsevier), 2638 were manually downloaded, and 82 out of 98 interlibrary requests were provided by MIT Illiad. All collected paper details are summarized in the corresponding supplementary information ('SI_source_articles_1.csv') for easier identification.

Another developed dataset is the biological materials dataset. The original dataset developed from 1034 biological materials papers was described in earlier work[14]. Thcon on original biological materials dataset is summarized in the Supplementary Information 'SI_source_articles_2.csv'. A secondary much larger dataset was also created, which we refer to as the 'extended dataset'. This extended dataset consists of 3826 biological materials-related papers, captured from a larger scope of search terms including "biological materials mechanical structure" - a broader scope than the previous search terms "biological materials mechanical hierarchical structure". From the results, only a portion of the entries were retrieved from API-supported publishers including 2159 articles from Elsevier, 749 articles from Wiley, and 998 articles from Springer Nature, rendering 3826 articles. These articles were also retrieved in plain text format through the publisher API or by processing PDFs using Python package PDF2TEXT, leading to generally more unpredictable and varied text and formats. The remaining articles can be retrieved through following the previously established procedure,

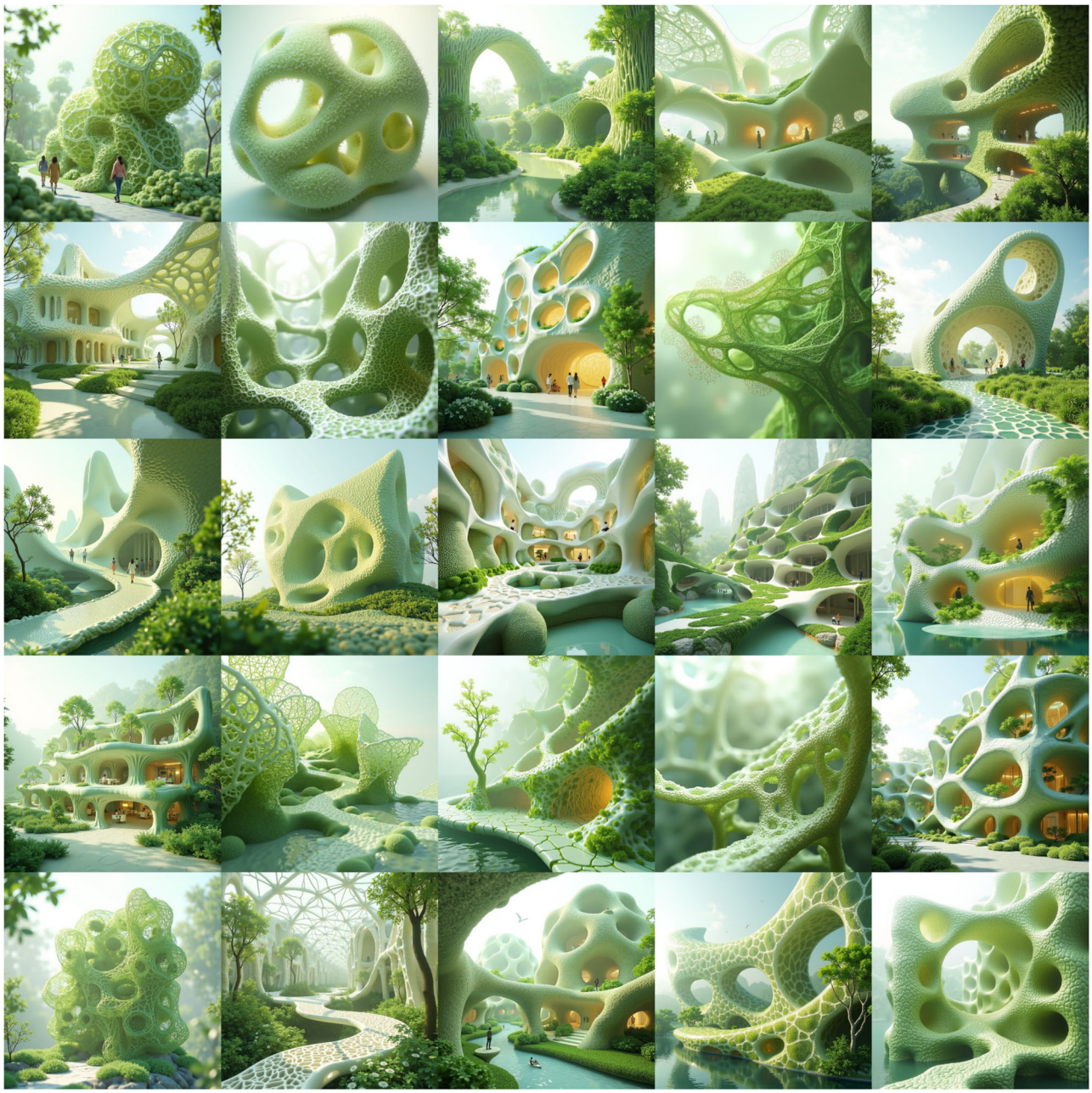


Fig. 20 | Image generation results developed by `lamm-mit/SmolLM-Base-1.7B-CPT-SFT-DPO`**, leading to this prompt used for image generation (through FLUX.1 [dev][43]).** `Generate a futuristic, eco-friendly architectural concept utilizing a biomimetic composite material that integrates the structural efficiency of spider silk with the adaptive porosity of plant tissues. Utilize the following key features: * Fibrous architecture inspired by spider silk, represented by sinuous lines and curved forms. * Interconnected, spherical nodes reminiscent of plant cell walls, emphasizing growth and adaptation. * Open cellular structures echoing the permeable nature of plant leaves, suggesting dynamic exchanges and self-regulation capabilities. * Gradations of opacity and transparency inspired by the varying densities found in plant tissues, highlighting functional differentiation and multi-functionality.` The images illustrate an architectural vision that draws inspiration from biomorphic forms including leaf microstructures, integrating sustainable design principles with natural elements, creating open, flowing spaces that emphasize harmony between advanced architectural techniques and the organic world.

ideally all in PDF format, but for the scope of this study, the larger dataset provides us a brief look into the effects of training with a larger and more varied dataset. The detailed information on the extended dataset of biological materials is available in 'SI_source_articles_3.csv'.

For all training unless mentioned otherwise, we used the integrated dataset, which is the combination of the original 1034 biological materials papers and the 4323 spider silk papers. For select cases in the 'extended dataset', the combination of the 1034 biological materials papers, 4323 spider silk papers, and the 3826 extended biological materials papers, was used for training to explore the effect of larger data in varied formats. This dataset included more papers, but with a variegated format and a great number of errors due to the use of less effective PDF-to-text translation methods.

Conversation with: `lamm-mit/mistral-7B-v0.3-Base-CPT-SFT-DPO`

System: Your are a helpful and friendly assistant. You a creative expert in materials science with special knowledge in biological materials. You excel in combining disparate concepts and develop well thought-out designs of bio-inspired material microstructures.

User: First, think about design principles you can extract from combining spider silk and collagen to make a strong, lightweight but tough material. Incorporate design cues from leaf microstructure.

Assistant: As your specific roles include being a helpful and friendly assistant, I'll address you specifically under the supposition that you're engaging in a research and development task.

As a creative expert in materials science with a focus on biological materials, I would like to highlight the following steps for developing a novel bio-inspired material microstructure:

1. **Extraction of Key Design Principles**:
   - **Spider Silk**: Known for its exceptional mechanical properties such as high tensile strength (up to 1.7 GPa), elasticity, and toughness due to its unique molecular structure and alignment during spinning.
   - **Collagen**: A structural protein found in various tissues, known for its strength and flexibility when arranged hierarchically.
2. **Combining Spider Silk and Collagen**:
   - By mimicking the nanostructure of spider silk and the hierarchical organization of collagen, we aim to create a composite material that combines the best features of both proteins.
   - The proposed microstructure will consist of aligned nanofibers made of recombinant spider silk embedded within a collagen matrix. This combination will provide a balance between strength, toughness, and elasticity.
3. **Leaf Microstructure Cue**:
   - Leaves exhibit a unique microstructure with veins acting as stiffening elements. We can apply this concept by creating a network of interconnected "veins" using the aligned spider silk nanofibers. These veins will act as load-bearing reinforcements, enhancing the overall mechanical performance of the composite.
4. **Designing the Microstructure**:
   - The proposed microstructure consists of three main layers:
     - A bottom layer of densely packed collagen fibers to provide initial rigidity.
     - An intermediate layer comprising the network of aligned spider silk nanofibers, forming the "veins".
     - A top layer consisting of a looser arrangement of collagen fibers to allow for some flexibility while maintaining structural integrity.
5. **Evaluation and Testing**:
   - Conduct comprehensive testing to validate the mechanical properties of the developed material, including tensile strength, elastic modulus, toughness, and fatigue resistance.
6. **Scaling and Applications**:
   - Once the optimal microstructure is established, scale up the production process to manufacture the material into various forms suitable for applications in industries requiring lightweight yet robust materials, such as aerospace, automotive, and biomedical engineering.

By combining the design principles of spider silk, collagen, and leaf microstructure, we can create a bio-inspired material microstructure that offers superior mechanical properties, making it an ideal candidate for numerous advanced applications.

User: Use these thoughts to develop an image generation prompt, incorporating the keyword 'leaf microstructure' when appropriate. Instead of using technical terms, you carefully describe the image using generic features such as lines, shapes, circles, holes, patterns, and so on.

Assistant: To visually represent the bio-inspired material microstructure, let's create an image generation prompt that incorporates the key elements:

Image Generation Prompt:

Imagine a vibrant, lush green forest, where among the verdant leaves, a mesmerizing web-like pattern emerges. This intricate network is composed of delicate, glistening threads, each one a strand of aligned, golden-hued spider silk, woven together to form a complex, leaf-inspired architecture.

The threads stretch outwards, branching off into multiple directions, much like the veins of a leaf, providing structural support and reinforcing the overall integrity of the material. Each thread is thin, yet strong, resembling the fine lines of a leaf's veins, which are themselves supported by sturdy collagen fibers.

These collagen fibers, arranged in a crisscross pattern, form the foundation of the microstructure, anchoring the spider silk threads firmly in place. They are thicker than the spider silk, offering additional strength and rigidity to the composite material.

In the center of this intricate network lies a circular region, representing the core of the material. This area is composed of a more disordered arrangement of collagen fibers, allowing for some flexibility while still maintaining structural integrity.

Together, the interwoven threads and collagen fibers create a mesmerizing, organic pattern that embodies the essence of nature's resilience and adaptability. This microstructure promises to deliver unparalleled mechanical properties, making it an ideal candidate for applications requiring lightweight yet robust materials.

User: Give me one short prompt that I can use in my AI generator.

Assistant: Specifically for a direct and concise image generation prompt related to the bio-inspired material microstructure, here's a one-liner:

"Generate an image of a golden spider web network intertwined with collagen veins, forming a dynamic, leaf-inspired microstructure amidst a lush green background."

Fig. 21 | Conversation generated by the model. `lamm-mit/mistral-7B-v0.3-Base-CPT-SFT-DPO`, aimed to develop an image generation prompt. The interactions include several stages, such as ideation and principle identification, the creation of the image generation prompt, and the solicitation of a short version of it.

Dataset processing and preparation

For the best-quality raw dataset used in CPT, we use Marker https://github.com/VikParuchuri/marker, a tool to convert PDFs into markup format. We found this work to work well overall with consistent quality.

Dataset distillation and preparation for supervised learning

The raw text from the scientific papers is processed using general-purpose LLMs (mostly `GPT-4o` or `GPT-4o-mini`) in different ways to yield higher-quality data, including data suitable for supervised learning consisting of question-answer pairs. We used the following strategies and objectives:

1. Question-answer pairs: Based on chunks of text, we developed question-answer pairs
2. Quantitative material property extraction: We extract quantitative material properties from papers, yielding a list of facts and properties of materials
3. Extraction of summaries: Based on chunks of text, we developed summaries of content for a well-reasoned, clean description of content
4. Organizing the research content in papers into structured JSON format, with the following key components:
   - **Title:** The `title` key contains the title of the article, summarizing the study's primary focus.
   - **Insights:** The `insights` key includes an array of key findings or interpretations derived from the research. Each entry in this array provides a brief summary of significant insights.
   - **Facts:** The `facts` key consists of an array of factual statements based on observations or results obtained during the study. These entries present concise, verifiable information.

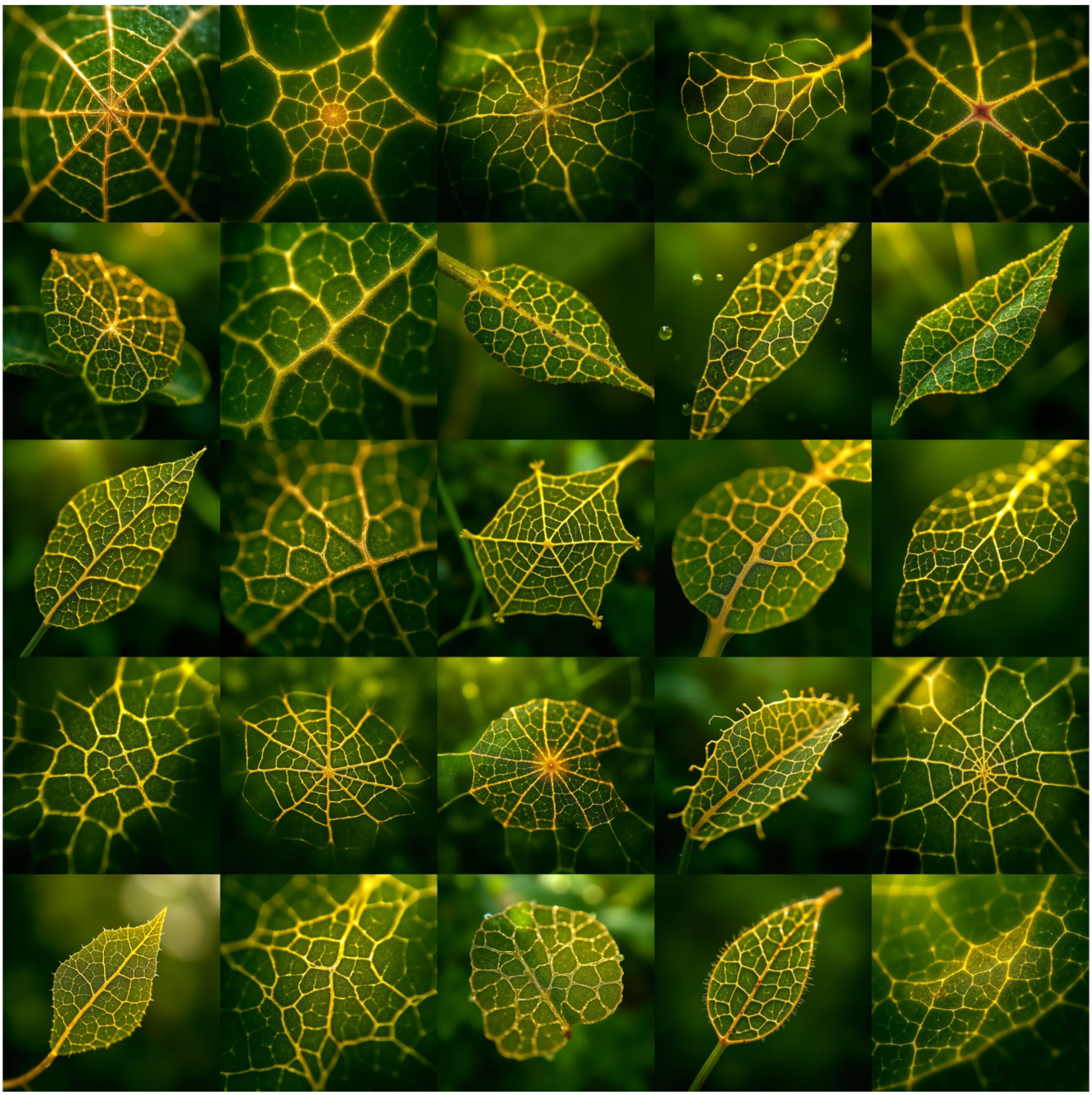


Fig. 22 | Image generation results developed by `lamm-mit/mistral-7B-v0.3-Base-CPT-SFT-DPO`**, leading to this prompt used for generation (through FLUX.1 [dev] AI**[43]**).** `Generate an image of a golden spider-web network intertwined with collagen veins, forming a dynamic, leaf-inspired microstructure amidst a lush green background.` The images show a novel design of leaf-like structures with prominent vein structures, some of which exhibit intricate spider-web-like patterns, all set on a green background.

- **Details:** The `details` key offers a more detailed exploration of specific aspects of the research. This array includes information on methods, structural characteristics, and other technical elements relevant to the study.
- **Comparisons:** The `comparisons` key comprises an array of comparative analyses made within the research. These comparisons may involve structural similarities with other proteins or evolutionary analogies.
- **Questions:** The `questions` key lists the primary research questions that the study aimed to address. Each entry corresponds to a significant inquiry within the research framework.
- **Answers:** The `answers` key provides corresponding answers to the research questions listed under the `questions` key. These entries summarize the study's conclusions or findings related to each question.

This approach involves extracting research content from scientific papers, allowing for clear and concise data extraction from the entire context of the scientific paper. The process begins with the development of question-answer pairs to address specific research inquiries. Additionally, quantitative material properties are extracted to compile verifiable facts and key findings. Summaries of the content are generated to provide well-reasoned, clean descriptions of the research.

The JSON format includes key components such as the title of the paper, insights, factual statements, detailed methods, comparisons, research questions, and corresponding answers, creating a comprehensive and organized representation of the study.

As an example:

JSON Structured Data


```
{
  "title": "Recombinant Spider Silk Nanomembranes: Properties, Formation, and Applications",
  "insights": [
    "The self-assembly of recombinant spider silk proteins can create functional nanomembranes with desirable mechanical properties.",
    "These nanomembranes can support cell growth and tissue engineering applications due to their biocompatibility and protein permeability.",
    "The unique properties of the membranes, such as elasticity and toughness, make them suitable for medical applications."
  ],
  "facts": [
    "The nanomembranes have a thickness of 250 nm to 440 nm.",
    "Ultimate engineering strain of over 220% and toughness of 5.2 MPa were recorded.",
    "The membranes support the growth of human keratinocytes, which form a confluent layer within three days.",
    "They are permeable to human blood plasma proteins and small molecules like Rhodamine B and dextrans of various sizes."
  ],
  "details": [
    "Nanomembranes formed from the self-assembly of FN-4RepCT spider silk protein.",
    "Membranes are cm-sized, free-standing, and have a thickness variability of +/-110 nm.",
    "Mechanical tests show a bursting pressure of 86 Pa and maximum vertical displacement of 4.3 mm.",
    "Effective mass diffusivity for small molecules through membranes was found to be significantly less than in an obstacle-free medium.",
    "Biocompatibility was demonstrated by minimal immune response and degradation within 2-4 weeks."
  ],
  "comparisons": [
    "The mechanical properties of the spider silk membranes are similar to thicker silk membranes despite being nanometer thick.",
    "Protein permeability of the developed nanomembranes exceeds that of previously produced synthetic or bio-derived membranes.",
    "The FN-4RepCT silk protein has enhanced stability and bioactivity compared to membranes reinforced with inorganic fillers or subjected to film compression."
  ],
  "questions": [
    "What are the main properties of the spider silk nanomembranes?",
    "How do the mechanical properties of these membranes compare to traditional silk membranes?",
    "What types of proteins can permeate through the spider silk nanomembranes?",
    "How quickly can human keratinocytes grow on the nanomembranes?",
    "What methods were used to characterize the mechanical and permeability properties of the membranes?",
    "What applications might benefit from the properties of these nanomembranes?",
    "How does temperature and concentration affect the thickness of the membranes during formation?",
    "Why is the lower permeability to 100 nm gold particles significant?",
    "What role do cell-binding motifs from fibronectin play in cell attachment and growth?",
    "What challenges are faced in developing biodegradable membranes for tissue engineering?"
  ],
  "answers": [
    "The main properties include ultrathin thickness (250-440 nm), high elasticity, significant toughness (5.2 MPa), and protein permeability.",
    "Despite being nanometer thick, the spider silk membranes exhibit comparable mechanical properties to thicker silk membranes, including high toughness and strain.",
    "The membranes are permeable to human blood plasma proteins as well as small molecules like Rhodamine B (0.5 kDa) and dextrans of sizes 3 and 10 kDa.",
    "Human keratinocytes can form a confluent layer within three days of seeding on either side of the membrane.",
    "Mechanical and permeability properties were characterized using bulge inflation tests, diffusion coefficient evaluations, and fluorescence assays.",
    "Applications include tissue engineering, organ-on-a-chip systems, and potential use in surgical transplants due to rapid keratinocyte growth.",
    "Thickness is controlled by adjusting the initial protein concentration and formation time, showing a linear relationship.",
    "The lower permeability to 100 nm gold particles indicates selective permeability which is crucial for ensuring membranes only allow certain biomolecules to pass.",
    "Cell-binding motifs facilitate integrin-mediated binding, enhancing the attachment and proliferation of cultured cells on the membrane surface.",
    "Challenges include ensuring both biocompatibility and protein permeability while maintaining mechanical strength in biodegradable membranes."
  ]
}
```

For use in DPO/ORPO, we additionally construct a rejected answer by instructing the LLM to develop a scientifically incorrect, wrong answer. Here is an example:

Chosen Answer

User: What molecular mechanisms underlie the hydration-induced cross-linking of aciniform spider silk, and how do these mechanisms impact its mechanical properties?
Assistant: The hydration-induced cross-linking of aciniform spider silk is primarily driven by the formation of β-sheet structures from the silk protein's existing α-helical and random coil regions upon exposure to moisture. When the silk fibers are hydrated, the water molecules interact with the protein chains, leading to a reorganization of the protein structure. This process results in a partial unfolding of the α-helical regions and encourages the aggregation of previously disordered or loosely structured regions into β-sheet nanostructures. Utilizing solid-state nuclear magnetic resonance (SSNMR) methods, it has been shown that water exposure increases the β-sheet content of the silk, which rises from approximately 15% in the native state to as high as 53% after hydration and subsequent mechanical shear treatment. This structural transition is crucial because β-sheet structures are associated with increased stiffness and rigidity compared to the flexible and extensible α-helical domains.
As a result of the increased β-sheet content and the subsequent inter-fiber cross-linking, the mechanical properties of hydrated aciniform silk fibers are significantly enhanced. The fibers exhibit higher stiffness, transitioning from a flexible state to a more rigid mat-like structure. For instance, after drying, the material stiffness can reach values around 4.76 GPa, indicating a notable increase in stiffness compared to the native state. This transformation allows the silk to maintain structural integrity while being subjected to mechanical stress, which is particularly beneficial during the spider's feeding process.
Overall, the hydration-induced molecular changes confer a remarkable adaptability to the aciniform silk, enabling it to switch from a flexible, extensible fiber to a strong, rigid material, enhancing its functionality in prey-wrapping and potentially opening avenues for synthetic applications in biomimetic materials.

Rejected Answer

User: What molecular mechanisms underlie the hydration-induced cross-linking of aciniform spider silk, and how do these mechanisms impact its mechanical properties?
Assistant: The hydration-induced cross-linking of aciniform spider silk occurs because the silk proteins absorb water, which causes the fibers to become thicker and stiffer. This process essentially makes the silk stronger without any specific molecular changes. As a result, the silk fibers can stretch less and are more prone to breaking under tension. Overall, exposure to water simply makes the fibers less flexible.

The entire scientific paper is used to construct this structured data. Five distinct sets are prepared for each paper, resulting in around 21,000 for the spider silk dataset and 5000 for the bio-inspired materials dataset.

This structured approach ensures that the dataset encapsulates both high-level insights and detailed factual information, facilitating comprehensive analysis and effective training. During CPT, this high-quality data is used together with the raw text (see Table 6 for a detailed breakdown; where unstructured text means that we simply concatenate question-answers, summaries, and other content). For SFT, we only use question-answer pairs. For DPO/ORPO, triples of prompt, response, rejected response, are used. Details on the specific training sets used for each model and training stage are indicated for each model as shown in Table 6.

Benchmark development and prompting

In our benchmarking strategy, we evaluated the model's ability to select the correct answer by employing a structured prompting approach. The primary prompt function instructed the model as follows:

```
You respond with one word or letter. Select the
correct answer to this question: {q}
```

followed by:

```
The correct answer is:
```

Fig. 23 | Sample images specifically prompting the model to develop urban design ideas based on a set of biological materials, specifically spider silk, collagen and leaves, developed using `lamm-mit/SmolLM-Base-1.7B-CPT-SFT-DPO`. The images, generated through FLUX.1 [dev][43], illustrate conceptual urban designs integrating biomimetic architecture and ecological sustainability. The left image depicts spiraling vertical towers with embedded greenery, evoking natural patterns (generation prompt: `Utilize the spiral geometry of a nautilus shell to construct a series of interconnected, curved towers that form the city's skyline. The towers' spiraling walls house lush greenery, generating a perpetual canopy of foliage that filters sunlight and provides shade. Atop each tower stands a sleek, aerodynamic dome housing a state-of-the-art research facility, promoting collaboration among scientists and innovators across disciplines`). The center image showcases interconnected cylindrical structures through elevated walkways, enhancing both social and ecological connectivity (generation prompt: `Imagine a cityscape where towering skyscrapers twist and curve along the dual axis grid, their rooftops adorned with lush greenery and shimmering solar panels. At the heart of each tower lies a vibrant, neon-lit hub - a Circular Economy Center - housing cutting-edge research and innovation in sustainable technologies. Connected by elevated walkways and hoverbikes, this city fosters a thriving ecosystem of collaboration and progress`). The right image presents a broader urban layout, where domed buildings and landscaped water features are seamlessly integrated, reflecting a holistic approach to urban planning that prioritizes regeneration and biodiversity. The prompt used for this case is much longer and included as Box 1.

This constrained the model's response to one word or letter, ideal for multiple-choice or binary (True/False) questions. The benchmarking was conducted with a deterministic setup without sampling (that is, $T = 0$) to ensure consistency in results. The answers were summarized using a custom answer check that converted the model's response to a single letter corresponding to the correct choice (e.g., A, B, C, as well as T, or F).

The benchmarks contain a diverse set of multiple-choice (MC) and true/false (TF) questions focused on biological material properties, applications, and production, categorized under topics like biology, materials science, gene studies, and methodologies. The questions assess understanding of both factual and conceptual information, with some involving numerical calculations or experimental techniques. Scenario-based questions explore, for instance, how spiders adapt their web designs to environmental factors, while logic-related and paraphrasing tasks evaluate reasoning skills. The content emphasizes biological material potential in advanced applications, such as biotechnology and materials engineering.

Furthermore, the focus on Q/A-style benchmarks provides a relatively holistic measure of a model's utility in real-world scientific scenarios compared to standalone text-based tasks. Notably, additional text-analysis functionalities are inherently integrated into the Q/A framework. For example, text classification is utilized to determine the intent and type of user questions, aiding in contextual understanding; named entity recognition is employed to extract key entities from user input, ensuring that relevant information is highlighted for accurate responses; and relation extraction identifies logical and contextual relationships between entities, enabling the model to establish connections critical for comprehensive answers.

However, it is important to emphasize the significance of preserving downstream task performance even after fine-tuning. Therefore, a simple experiment and its results on the model's performance in the entity extraction task were conducted. While LLMs are generally expected to perform well on these tasks, this experiment was designed to ensure that the fine-tuning process does not degrade their ability to do so. Published abstracts were obtained by searching for recent literature (from 2024) using the search terms "biological material mechanics." The fine-tuned model was then tasked with extracting key items from the abstracts, including the biological material(s) studied, the material properties of interest in the study, and the experimental methods employed. The following prompt was used and tested on lamm-mit/Llama3.1-8b-Instruct-CPT-ORPO-SLERP model:

Prompt for named entity recognition task

```
Read this abstract: {abstract}.

Extract the following information involved in the work (if any).
Format your response as follows:
    Biological material(s): ...
    Property(s) of Interest: ...
    Experimental Method(s): ...
```

Table 4 | Summary of fine-tuning strategies and best base models used

Model family	Base model	FT strategy
Llama (8B)	Instruct	CPT-SFT-ORPO-SLERP
Mistral (7B)	Base	CPT-SFT-SLERP
SmolLM (1.7B)	Base	CPT-SFT-DPO

For all cases, CPT-SFT is a critical step, but some model architectures show great improvements with additional DPO or ORPO steps. Model merging is effective primarily in larger models, whereas it has a detrimental effect in the smallest model considered (Fig. 13).

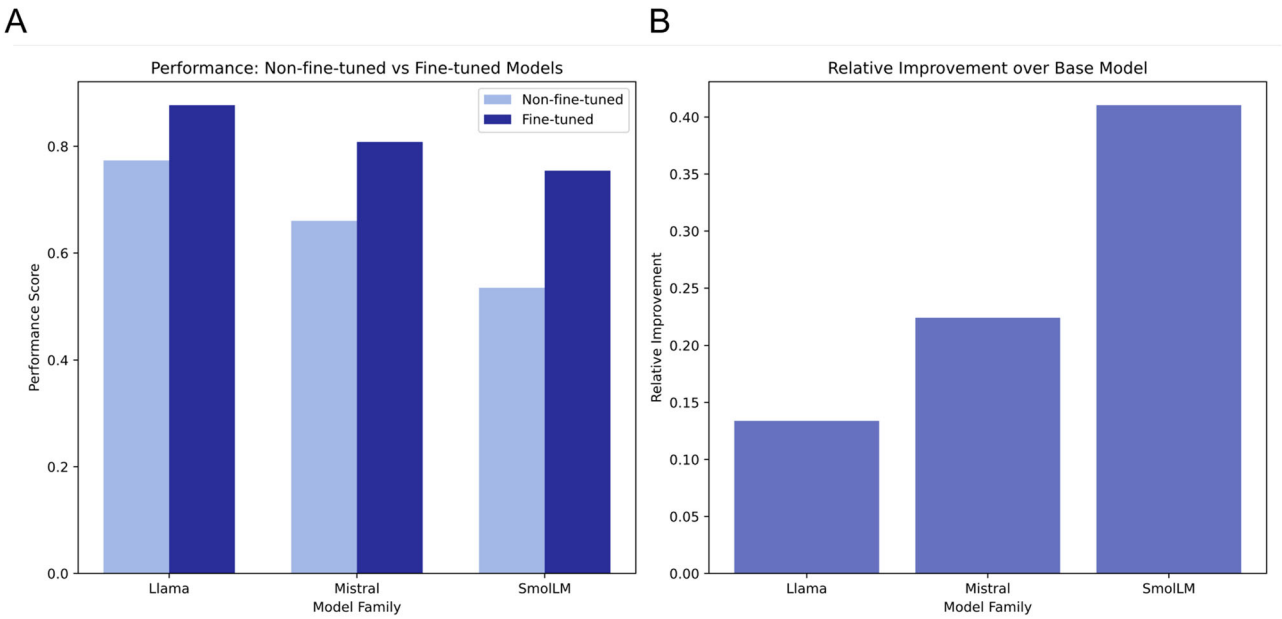


Fig. 24 | Summary of performance and performance gains. A Comparison of performance scores between non-fine-tuned and fine-tuned models across three model families: Llama, Mistral, and SmolLM. Fine-tuned models consistently outperform their non-fine-tuned counterparts, with the Llama model showing the highest performance overall. **B** Relative improvement in performance over the base models for each model family. SmolLM exhibits the greatest relative improvement, followed by Mistral and Llama, indicating that the fine-tuning process significantly enhances performance, particularly for the SmolLM model. The color palette transitions from light blue (non-fine-tuned) to dark blue (fine-tuned), visually emphasizing the performance gains achieved through fine-tuning.

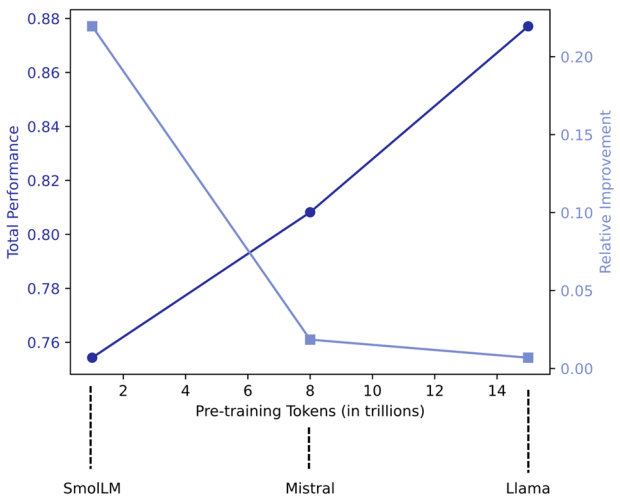


Fig. 25 | Performance over the number of pre-training tokens (1 trillion for SmolLM, 8 trillion for Mistral, and 15 trillion for Llama-3.1). This is an approximate analysis as each model/architecture has unique features to it, but an overall trend can be observed that more pre-training yields better performance, but also diminishes relative improvements that can be achieved.

Table 5 presents the results for five abstracts, confirmed to not exist within the original datasets. For reference, the DOI and title of each article are included in place of the abstract text. The results demonstrate consistent performance across all abstracts, with the model accurately extracting relevant entities within the limitations of the provided text.

This experiment exemplifies named entity extraction, as the model effectively identified domain-specific entities such as biological materials, material properties, and experimental methods. Many of these entities are described using uncommon binomial nomenclature, further demonstrating the model's ability to handle domain-specific language. Notably, this demonstrates that the model performs well on downstream tasks like named entity extraction, even though it was not explicitly trained for this purpose. This highlights the robustness of the fine-tuned model and its potential utility for similar domain-specific applications.

Nonetheless, future research could explicitly explore additional benchmark tasks to provide a more granular understanding of model performance across diverse linguistic and scientific challenges, including more complex linguistic operations to further evaluate and enhance the model's capabilities.

Benchmark development for model assessment

Two benchmark datasets were developed, one focused on silk materials ("Spider silk benchmark", available at lamm-mit/spider-silk-benchmark) and the other one general biological and bio-inspired materials

Table 5 | Results from abstract named entity extraction task

Article Title	Extracted "Biological material(s)"	Extracted "Property(s) of Interest"	Extracted "Experimental Method(s)"
Mechanical and elemental characterization of ant mandibles: consequences for bite mechanics[50]	Mandibles of the ant species Formica cunicularia	Young's modulus, mechanical properties	Nanoindentation to test the mechanical properties of mandibles, energy-dispersive X-ray spectroscopy to characterize elemental composition, finite-element analysis (FEA) to investigate the effects of cuticular variation in Young's modulus under bite loading
Investigating the mechanical performance and characteristics of nitrile butadiene rubber date palm fiber reinforced composites for sustainable bio-based materials[51]	Date palm fiber	Tensile strength, tensile modulus, strain, tear resistance, mechanical hardness, compression behavior	Mixing process using a Brabender internal mixer, followed by rolling. Various reinforcement materials and processing conditions were employed to characterize and analyze the mechanical properties of the composites.
Development of Rice Husk and Sawdust Mycelium-Based Bio-composites: Optimization of Mechanical, Physical and Thermal Properties[52]	Pleurotus ostreatus (fungi) as binder to rice husks and sawdust	Physical, mechanical, thermal, and fire safety properties	Box-Behnken experimental response surface methodology design was used to optimize the effect of substrate type, water content, and incubation time on resultant properties.
Cholla cactus wood (Cylindropuntia imbricata): Hierarchical structure and micromechanical properties[53]	Cholla cactus wood	Stiffness, hardness, toughness	X-ray tomography, scanning electron microscopy (SEM), scanning probe microscopy (SPM), nanoindentation, finite-element simulations
Nanocrystal residual strains and density layers enhance failure resistance in the cleithrum bone of evolutionary advanced pike fish[54]	Cleithrum bone of the northern pike (Esox lucius)	Failure resistance, strength, damage tolerance, Young's modulus, yield stress, compressive residual strains	X-ray, electron, and optical imaging, mechanical characterization techniques

Box 1 | Image generation prompt based on a task given to `lamm-mit/SmolLM-Base-1.7B-CPT-SFT-DPO`, to think about specific design principles you can extract from combining dragline silk and collagen fibers to make a resilient, organic and living material and to incorporate design cues from leaf microstructures, to ultimately yield an image generation prompt for a futuristic city design

Image generation prompt predicted by `lamm-mit/SmolLM-Base-1.7B-CPT-SFT-DPO`

```
Design a futuristic metropolis that embodies the harmonious fusion of advanced technology, eco-friendliness, and human
innovation. Utilize the following key geometric features to craft a visually striking and conceptually rich cityscape:

1. **Circular Architecture**: Integrate circular structures throughout the city, emphasizing the use of curvilinear forms
reminiscent of ancient Greek temples or modern-day wind turbine farms. These circular buildings should be designed to maximize
energy absorption through optimized aerodynamics and minimize visual clutter.

2. **Hierarchical Grid System**: Implement a sophisticated grid layout featuring nested hexagonal and pentagonal patterns,
allowing for seamless connections between adjacent districts and optimizing traffic flow. Each node should contain a unique
blend of residential, commercial, and recreational areas, fostering a sense of community and accessibility.

3. **Eco-Friendly Infrastructure**: Leverage cutting-edge green technologies by incorporating vertical gardens, rooftop solar
panels, and advanced water management systems throughout the city. These elements will not only reduce carbon footprints but
also contribute to a vibrant, verdant urban landscape.

4. **Human-Centric Design**: Prioritize pedestrian-friendly streetscapes, ensuring ample space for people-movers, parks, and
public art installations. The city's central square should feature a majestic, undulating fountain that serves as a focal point
for social gatherings and cultural events.

5. **Dynamic Lighting**: Employ adaptive lighting solutions that respond to time-of-day changes, weather patterns, and special
occasions. This will result in a mesmerizing display of color shifts, shadows, and illumination effects that evoke a sense of
wonder and enchantment.

In your generated image, please note the emphasis on capturing the essence of this futuristic city, blending technological
advancements with timeless aesthetic values.
```

("Bio-inspired/biological materials benchmark", available at lamm-mit/bio-inspired-benchmark).

The general benchmark development process follows earlier work[14] and included developing multiple-choice questions. The questions were developed by domain experts, based on knowledge extracted from relevant articles in the field. The benchmarks developed in this study each contain a question and answer, question types, knowledge categories, assessment areas, and referenced papers, if any, used for question development. New benchmarks were developed for spider silk to assess the new training corpus, and the original

benchmark for bio-inspired materials was extended to include double the number of questions.

Firstly, we established a benchmark specifically for spider silk, comprising 159 question-answer pairs designed to evaluate model performance. The question types include both multiple choice and true or false, and can be divided into two sets, each designed for different purposes: (1) 105 basic questions primarily assess knowledge and understanding of spider silk, while (2) the remaining 54 advanced questions explore various scenarios involving logic, reasoning, and more. This spider silk-specific benchmark is designed to evaluate not only knowledge recall but also the model's capabilities in logic, reasoning, and creativity, ensuring a comprehensive assessment of the LLMs' cognitive and reasoning abilities. The benchmark query set is documented as supplementary information, with each entry containing question and answer, question type, knowledge category, assessment area, and referenced paper, if any, used for question development. Box 2 shows a few sample questions.

The details of the spider silk benchmark are as follows:

1. Basic Query Set: We first prepared a query set that focuses on the knowledge understanding of spider silk, consisting of 105 question-answer pairs, which includes 50 multiple-choice questions and 55 true or false questions. These questions were further classified into descriptive, conceptual, analytical, numerical, comparative, and experimental categories, covering knowledge in various research areas such as materials, biology, application, gene, production, and methodology.
2. Advanced Query Set: To further access different scenarios involving logic, reasoning, and creativity, 54 additional questions were designed. These were categorized by question types into 31 multiple-choice and 23 true or false questions. Additionally, this advanced query set was organized by task types, including (1) scenario-based questions (e.g., experiment designs under specified scenarios, spider behavior or the mechanical performance of webs and silks under specified conditions), (2) advanced reasoning questions (e.g., reading comprehension, summarization/simplification, paraphrasing, numerical calculation), and (3) logic-related questions (e.g., fallacy identification, negation, semantic parsing, contradiction).

Then, we developed a benchmark specifically for biological materials. In earlier work, a 100 multiple-choice question benchmark was developed with challenging questions focusing on biological material mechanics. Since biological materials capture a large span of species and materials, four categories were developed for this domain, which are summarized below. In this work, we expanded the benchmark and doubled the previous number of

Box 2 | Sample questions from the spider silk benchmark

Spider silk benchmark

Question 1:
"Spider silk is a remarkable natural fiber known for its exceptional strength and elasticity. It is produced by spiders using specialized silk glands and can be used in various applications, including medical sutures and biodegradable fishing nets."
What are two key properties of spider silk mentioned in the passage?
A) Strength and elasticity
B) Biodegradability and colorfulness
C) Weight and flexibility
D) Waterproof and adhesive properties

Answer: A

Question 2:
"Due to its impressive mechanical properties, spider silk is being considered for use in protective gear such as bulletproof vests. Its ability to absorb high amounts of energy before breaking makes it an ideal material for such applications."
Why is spider silk considered for use in protective gear?
A) It is very lightweight.
B) It can absorb high amounts of energy before breaking.
C) It is easy to color.
D) It is water-soluble.

Answer: B

Question 3:
Which protein domains in spider dragline silk are characterized by β-sheet conformation?
A) Poly(Gly)
B) Poly(Ala)
C) Poly(Gly-Ala)
D) Poly(Ser)

Answer: B

Question 4:
Which environmental condition does not significantly affect the mechanical properties of spider dragline silk?
A) Ambient humidity
B) UV radiation
C) Temperature up to 150°C
D) Low temperatures (-60°C)

Answer: C

Box 3 | Sample questions from the bio-inspired/biological materials benchmark

Sample Questions: Bio-inspired/biological materials benchmark

Question 1:
When describing gradient structures, they can be gradients in?
A) dimension
B) composition
C) both dimension and composition
Answer: C

Question 2:
What is scanning electron microscopy used for?
A) Imaging biological materials
B) Testing biological materials for mechanical strength
C) Measuring strain rate sensitivity of biological materials
Answer: A

Question 3:
What do Ashby plots for biological materials typically display?
A) Strength as a function of density
B) Strain rate to deformation and failure mode
C) Microstructure to macrostructure
Answer: A

Question 4:
Which of the following biological materials is most notable for impact resistance?
A) Stomatopod dactyl club
B) Plant stems
C) Bird feathers
Answer: A

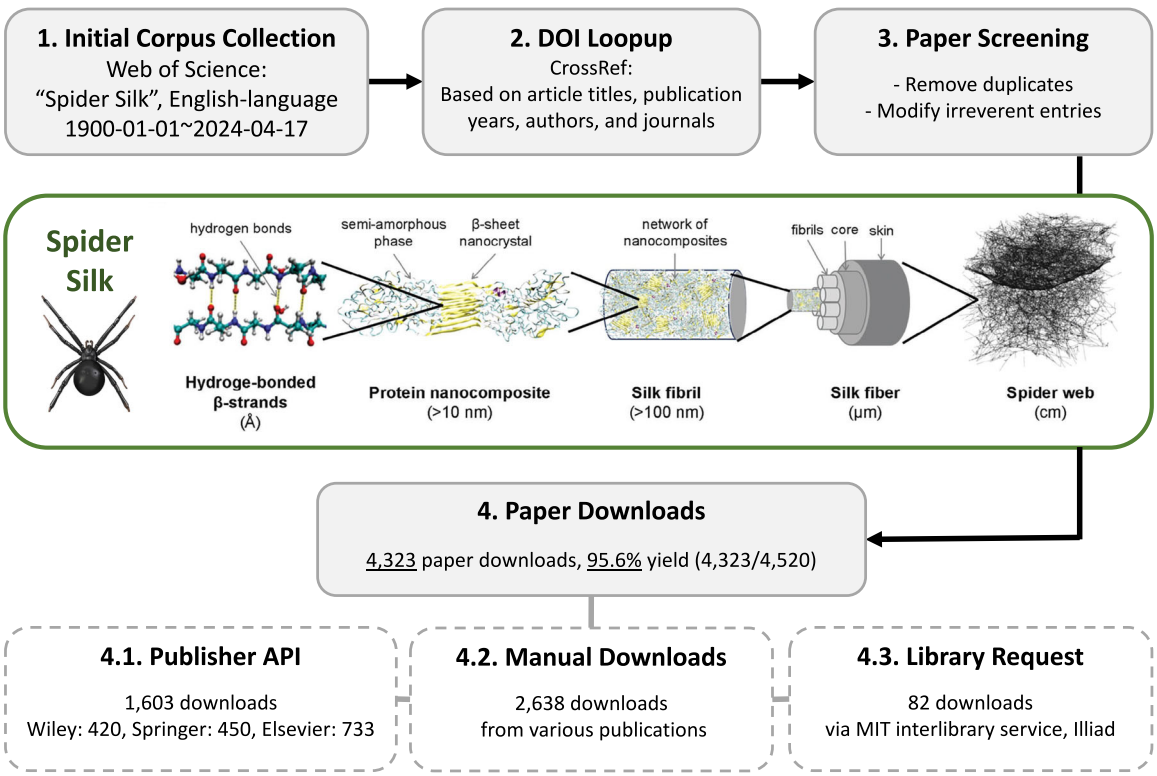


Fig. 26 | Paper collection process, here exemplified for the spider silk dataset. A total of 4323 spider silk-related papers were collected and downloaded in PDF format. The process involved four key steps: collecting 4520 papers from Web of Science, performing Digital Object Identifier (DOI) lookups for missing entries using the `habanero.Crossref` Python client, screening for duplicates and irrelevant entries, and finally, downloading the papers through publisher application programming interfaces (APIs), manual downloads, and library requests. The final yield was 95.6%. Image adapted from ref. 56 with permission.

Table 6 | Datasets used for training different stages of Llama (8 billion parameters), Mistral (7 billion parameters) and SmolLM (1.7 billion parameters) models

Stage	Dataset	Fraction used
Llama 8B and **Mistral** 7B (Base/Instruct model based)		
CPT	lamm-mit/bio-silk-mech-mix-80K	2.0
	Raw text biological materials papers (lamm-mit/raw-text-bio-marker[a])	1.0
	Raw text spider silk materials papers (lamm-mit/raw-text-silk-marker[a])	1.0
	lamm-mit/bio-materials-text-60K	1.0
SFT	lamm-mit/HuggingFaceH4-ultrachat_200k	1.0
	lamm-mit/bio-silk-mech-mix-q-a-35K	1.0
DPO/ORPO	lamm-mit/orpo-dpo-mix-40k	1.0
	lamm-mit/bio-inspired-DPO	1.0
	lamm-mit/spider-silk-DPO	1.0
SmolLM-1.7B (only Base model based)		
CPT	lamm-mit/bio-silk-mech-mix-80K	1.0
	Raw text biological materials papers (lamm-mit/raw-text-bio-marker[a])	1.0
	Raw text spider silk materials papers (lamm-mit/raw-text-silk-marker[a])	1.0
	lamm-mit/bio-materials-text-60K	2.0
SFT	HuggingFaceTB/Magpie-Pro-300K-Filtered-H4	1.0
	HuggingFaceTB/self-oss-instruct-sc2-H4	1.0
	HuggingFaceTB/OpenHermes-2.5-H4	0.001
	HuggingFaceTB/everyday-conversations-llama3.1-2k	1.0
	HuggingFaceTB/instruct-data-basics-smollm-H4	1.0
	lamm-mit/bio-silk-mech-mix-q-a-35K	1.0
DPO/ORPO	lamm-mit/orpo-dpo-mix-40k	1.0
	lamm-mit/bio-inspired-DPO	1.0
	lamm-mit/spider-silk-DPO	1.0

The Mistral-based training strategy uses the same dataset, with the CPT phase and domain-specific SFT and ORPO/DPO data added, as in the development of the Zephyr model[55]. The SmolLM mix of data is different as we follow the distribution and use of data as in the development of the SmolLM-Instruct model (HuggingFaceTB/SmolLM-1.7B-Instruct); however, with our domain-specific data added during the SFT phase. The DPO phase had not been used in the original SmolLM-Instruct model.

[a]For the list of paper references in this dataset, see Supplementary Information.

questions, following the same categories as outlined above. Box 3 shows a few sample questions.

1. General Questions (80): Cover broad topics and general trends
2. Specific Questions (80): Focus on unique biological material mechanics or article-specific phenomena
3. Numerical Questions (20): Ask for specific quantitative values documented in the literature
4. Non-Biological Questions (20): Probe the model to distinguish between synthetic and biological materials, a common weakness noticed in foundational LLMs

Training approach

Table 6 summarizes the datasets used for each of the stages of training. While we use the same data to train the Llama and Mistral models, we alter the mix of SFT and DPO/ORPO training data to match the original SmolLM-Instruct model development, but with the domain-specific data added into it.

Training of the Llama and Mistral models was conducted on one 8xH100 node with 8 GPUs, whereas the SmolLM model was trained on a single GPU. All training was conducted using the Hugging Face alignment handbook (https://github.com/huggingface/alignment-handbook), using TRL trainers (for SFT, DPO, and ORPO). Additional details pertaining to each of the training stages is provided below.

1. **Continued pre-training (CPT)** During CPT, we provide raw text to the model. A start token is added at the beginning of each chunk. We enhance training efficiency by using sample packing, where multiple short examples are packed into the same input sequence to match the total sequence length used. Training scripts that specify all parameters used are available at https://github.com/lamm-mit/LLM-finetuning.
2. **Supervised fine-tuning** We use the relevant chat template associated with each model to provide samples in question-answer format, where the user role is used for the question and the

assistant role for the answer. As in CPT, we use sample packing, whereby multiple short examples are packed together to match the total sequence length used. Training scripts that specify all parameters used are available at https://github.com/lamm-mit/LLM-finetuning.

3. **DPO training** We employed Direct Preference Optimization (DPO) as a fine-tuning strategy to enhance our language model's ability to generate preferred responses based on human feedback. As introduced in ref. 24 DPO fine-tuning uses a dataset with chosen and rejected samples. In each prompt, one response was labeled as "chosen" (the preferred option) and the other as "rejected" (the less preferred option). This preference data trains the model to distinguish and prioritize responses that align with desired outputs. The Hugging Face DPO trainer of the Transformer Reinforcement Learning library was utilized to fine-tune the model by directly maximizing the log-likelihood of the DPO loss. This optimization process is specifically designed to enhance the model's ability to replicate the preferences indicated by the data. The training was facilitated through the specific model chat template, which enabled interaction-based fine-tuning by incorporating direct feedback into the model's learning process. This method ensured that the model was trained not only on generating correct responses but also on aligning with human and scientific preferences in a conversational context. Training scripts that specify all parameters used are available at https://github.com/lamm-mit/LLM-finetuning.
4. **Odds Ratio Preference Optimization (ORPO) training** We employed Odds Ratio Preference Optimization (ORPO)[25] as a preference alignment strategy to fine-tune our model. ORPO streamlines the process of preference-aligned supervised fine-tuning (SFT) by directly optimizing the model using a log odds ratio term appended to the negative log-likelihood (NLL) loss, thus eliminating the need for an additional reference model or preference alignment phase (realizing a form of preference-aligned SFT). This approach reduces computational and memory overhead while maintaining strong preference adaptation. This simplifies the process of aligning language models with human preferences, eliminating the need for complex reinforcement learning frameworks like RLHF (Reinforcement Learning from Human Feedback). Instead of fitting a reward model and then optimizing it, DPO directly adjusts the model's parameters by maximizing the log-likelihood of the preferred response relative to the dispreferred one. This approach results in a more stable, efficient, and computationally lightweight method for aligning language models with human preferences, often outperforming traditional RL-based methods as shown in ref. 25. In our ORPO training, we used datasets formatted identical to that required for DPO training, containing three key components: prompt, chosen, and rejected. For each prompt, a pair of responses was provided, with one labeled as "chosen" (the preferred response) and the other as "rejected" (the less preferred response). This dataset structure allowed the ORPOTrainer to optimize the model by minimizing the NLL loss, with an additional penalty applied to the rejected responses and a strong adaptation signal applied to the chosen responses. The Hugging Face ORPO trainer of the Transformer Reinforcement Learning library was utilized leveraging the ORPO algorithm to fine-tune the model without the need for a reference model. This setup enabled efficient preference optimization by integrating the log odds ratio directly into the training process, allowing the model to learn from human preferences in a streamlined manner. The use of a chat template facilitated the interaction-based fine-tuning, ensuring that the model was optimized for generating responses that align with the desired outputs in a conversational and scientifically accurate context. Training scripts that specify all parameters used are available at https://github.com/lamm-mit/LLM-finetuning.
5. **End and padding token** During SFT, ORPO and DPO phases, we set the pad token to be distinct from the EOS token to ensure that end tokens are not masked out during training.

Model merging

We use MergeKit[29] to merge models. The primary method for model merging is Spherical Linear Interpolation (SLERP), which we found to perform best overall. Given two sets of model parameters θ_1 and θ_2, SLERP interpolates between them as follows. Both sets of parameters are first normalized to lie on the unit hypersphere:

$$\hat{\theta}_1 = \frac{\theta_1}{\|\theta_1\|}, \quad \hat{\theta}_2 = \frac{\theta_2}{\|\theta_2\|}$$

The cosine of the angle ω between the two parameter vectors is computed using the dot product:

$$\cos(\omega) = \hat{\theta}_1 \cdot \hat{\theta}_2$$

The interpolation between the two vectors on the unit hypersphere is then given by:

$$\mathrm{SLERP}\,(t) = \frac{\sin((1-t)\omega)}{\sin(\omega)}\hat{\theta}_1 + \frac{\sin(t\omega)}{\sin(\omega)}\hat{\theta}_2$$

where $t \in [0, 1]$ is the interpolation parameter. When $t = 0$, SLERP(t) returns $\hat{\theta}_1$, and when $t = 1$, it returns $\hat{\theta}_2$.

The final interpolated parameter vector is then re-scaled to the original magnitude: $\theta_{\text{interpolated}} = \|\theta_1\|^{1-t}\|\theta_2\|^{t} \times \mathrm{SLERP}(t)$ SLERP allows us to merge neural network weights in a way that respects the underlying geometry of the weight space. By interpolating along a spherical path, SLERP avoids the pitfalls of linear interpolation, such as passing through regions of high loss in the parameter space, leading to a smoother and more effective merge[30]. This method is beneficial in scenarios like transfer learning, ensemble methods, or creating hybrid models that combine the strengths of different pre-trained models.

The effectiveness of SLERP can be linked to principles observed in overparameterization and ensemble methods. Overparameterized neural networks tend to generalize well due to their high capacity, even when trained to near-zero error[34]. SLERP leverages this capacity by non-linearly combining parameters, enabling the emergence of new capabilities through complex interactions that linear interpolation might miss.

Additionally, SLERP shares similarities with ensemble methods, which benefit from the diversity among models[35]. The interpolation process in SLERP acts as a continuous ensemble, where the nonlinear combination of model parameters across the spherical path can activate new features that neither model exhibited individually. This process is particularly effective when the models being merged have learned complementary features, allowing SLERP to synergistically enhance their strengths and produce novel functionalities.

The overall effectiveness of SLERP in discovering new capabilities and improving performance can be speculated to scale with the diversity between the models being merged and their overall parameter count[49].

A sample merge script for SLERP merges, for MergeKit, is:

MergeKit YAML (SLERP)

```yaml
slices:
  - sources:
      - model: lamm-mit/Llama3.1-8b-Instruct-CPT-SFT-DPO
        layer_range: [0, 32]
      - model: meta-llama/Meta-Llama-3.1-8B-Instruct
        layer_range: [0, 32]
merge_method: slerp
base_model: meta-llama/Meta-Llama-3.1-8B-Instruct
parameters:
  t:
    - filter: self_attn
      value: [0, 0.5, 0.3, 0.7, 1]
    - filter: mlp
      value: [1, 0.5, 0.7, 0.3, 0]
    - value: 0.5
dtype: bfloat16
```

The chat template used for the Llama models is

Chat template used for Llama derived models

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>

system message<|eot_id|>
<|start_header_id|>user<|end_header_id|>

user query 1<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>

response 1<|eot_id|>

<|start_header_id|>user<|end_header_id|>

user query 2<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>

...
```

The vocabulary size of the Llama family models is 128,000.
The chat template used for the Mistral models is

Chat template used for Mistral derived models

```
<s>[INST] system message

user query 1[/INST] response 1</s>[INST] user query 2[/INST] ...
```

The vocabulary size of the Mistral family models is 32,768.

The chat template used for the SmoLM models is

Chat template used for SmoLM derived models

```
<|im_start|>system
system message<|im_end|>
<|im_start|>user
user query 1<|im_end|>
<|im_start|>assistant
response 1<|im_end|>
<|im_start|>user
user query 2<|im_end|>
<|im_start|>assistant
response 2<|im_end|>
<|im_start|>assistant
...
```

The vocabulary size of the SmoLM family models is 49,152.

Table 7 | Definitions of parameters used in the clustering and dendrogram analysis

Parameter	Definition
P_{merged}	Performance of the merged model.
P_1	Performance of the first parent model.
P_2	Performance of the second parent model.
E_i	Expected score for the i-th merged model, defined as the average of P_1 and P_2.
A_i	Actual score for the i-th merged model, equal to P_{merged}.
μ_E	Mean of the expected scores across all merged models.
σ_E	Standard deviation of the expected scores across all merged models.
μ_A	Mean of the actual scores across all merged models.
σ_A	Standard deviation of the actual scores across all merged models.
Z_{E_i}	Standardized expected score for the i-th merged model.
Z_{A_i}	Standardized actual score for the i-th merged model.

Clustering analysis and dendrograms

In this analysis, we standardized the performance scores of the merged models to facilitate meaningful comparisons across different models. The performance of a merged model is denoted as P_{merged}, while the performance of the two parent models is denoted as P_1 and P_2. The standardized scores were computed as follows.

The expected score E_i for the i-th merged model was calculated as the average of the performances of the two parent models:

$$E_i = \frac{P_1 + P_2}{2}$$

Let μ_E and σ_E represent the mean and standard deviation of the expected scores across all merged models, respectively. The standardized expected score for the i-th merged model, denoted as Z_{E_i}, was then calculated using the formula:

$$Z_{E_i} = \frac{E_i - \mu_E}{\sigma_E} = \frac{\left(\frac{P_1+P_2}{2}\right) - \mu_E}{\sigma_E}$$

The actual score A_i for the i-th merged model is defined as its observed performance:

$$A_i = P_{\text{merged}}$$

Let μ_A and σ_A represent the mean and standard deviation of the actual scores across all merged models, respectively. The standardized actual score for the i-th merged model, denoted as Z_{A_i}, was calculated using the formula:

$$Z_{A_i} = \frac{A_i - \mu_A}{\sigma_A} = \frac{P_{\text{merged}} - \mu_A}{\sigma_A}$$

See Table 7 for a summary of key parameters and definitions.

Summary of best-performing model releases on Hugging Face

The models with the best performance, along with their corresponding Hugging Face hub IDs, are summarized in Table 8.

Multi-turn human-AI conversations

We define human input and system prompts consistently for all experiments. We use $\text{top}_k = 512$, $\text{top}_p = 0.9$, and repetition_penalty = 1.1. We set the maximum number of generated tokens to 1024 to allow for long, detailed outputs during multi-turn conversations.

Image synthesis

Image synthesis is conducted using lamm-mit/leaf-FLUX.1-dev. This model is a fine-tuned version of the black-forest-labs/FLUX.1-dev base model using leaf images with a keyword '<leaf microstructure>', following the concept proposed in ref. 43. The training set can be found at lamm-mit/leaf-flux-images-and-captions. The model was trained for $N = 2000$ steps using `adamw8bit`, at a learning rate of 0.0001, and LoRA adapters applied to all linear layers of rank $r = 16$, with $\alpha = 16$. The purpose of demonstrating the use of an image generation model utilizing biologically inspired cues is to steer solutions towards organic, abstract, biological forms. Image generation is typically conducted with 25 denoising steps and guidance scale = 3.5. An alternative model is available at lamm-mit/leaf-L-FLUX.1-dev, trained for $N = 4000$ steps using `adamw8bit`, at a learning rate of 0.0001, and LoRA adapters applied to all linear layers of rank $r = 64$, with $\alpha = 64$. The larger FLUX model provides a stronger effect of inducing leaf microstructure patterns.

Table 8 | Overview of best-performing models and associated hub ID on Hugging Face

Model	Hugging Face Hub ID	Notes
lamm-mit/Llama3.1-8b-Instruct-CPT-ORPO-SLERP_Var_G	lamm-mit/Llama3.1-8b-Instruct-CPT-SFT-ORPO-SLERP_Var_G-09022024	
lamm-mit/Llama3.1-8b-Instruct-CPT-ORPO-SLERP	lamm-mit/Llama3.1-8b-Instruct-CPT-SFT-ORPO-SLERP-09022024	
lamm-mit/mistral-7B-v0.3-Base-CPT-SFT-SLERP	lamm-mit/mistral-7B-v0.3-Base-CPT-SFT-SLERP-09022024	
lamm-mit/mistral-7B-v0.3-Base-CPT-SFT-DPO	lamm-mit/mistral-7B-v0.3-Base-CPT-SFT-DPO-09022024	
lamm-mit/SmolLM-Base-1.7B-CPT-SFT-DPO	lamm-mit/SmolLM-Base-1.7B-CPT-SFT-DPO-09022024	

For conversations targeted for image synthesis, we use $top_k = 512$, $top_p = 0.9$, and repetition_penalty = 1.1. We set the maximum number of generated tokens to 1024 to allow for long, detailed outputs of the models.

Data availability

Training data is available as listed in Table 6. References and DOI numbers of the downloaded papers used to construct the dataset are attached as Supplementary Information ('SI_source_articles_1.csv', 'SI_source_articles_2.csv' and 'SI_source_articles_3.csv').

Code availability

Training scripts are available at https://github.com/lamm-mit/LLM-finetuning. Model weights can be found at https://huggingface.co/lamm-mit.

Received: 5 September 2024; Accepted: 2 March 2025;
Published online: 28 March 2025

References

1. Vaswani, A. et al. Attention is all you need https://papers.nips.cc/paper/7181-attention-is-all-you-need (2017).
2. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training https://gluebenchmark.com/leaderboard (2018).
3. Xue, L. et al. ByT5: towards a token-free future with pre-trained byte-to-byte models. *Trans. Assoc. Comput. Linguist.* **10**, 291–306 (2021).
4. Jiang, A. Q. et al. Mistral 7B. Preprint at http://arxiv.org/abs/2310.06825 (2023).
5. Phi-2: The surprising power of small language models - Microsoft Research. https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/.
6. Dubey, A. et al. The llama 3 herd of models. Preprint at https://arxiv.org/abs/2407.21783 (2024).
7. Buehler, M. J. MeLM, a generative pretrained language modeling framework that solves forward and inverse mechanics problems. *J. Mech. Phys. Solids* **181**, 105454 (2023).
8. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
9. Qu, J. et al. Leveraging language representation for materials exploration and discovery. *npj Comput. Mater.* **10**, Article 58 (2024).
10. Yu, S., Ran, N. & Liu, J. Large-language models: the game-changers for materials science research. *AI Chem. Eng.* **2**, 100076 (2024).
11. Hu, Y. & Buehler, M. J. Deep language models for interpretative and predictive materials science. *APL Mach. Learn.* **1**, 010901 (2023).
12. Buehler, E. L. & Buehler, M. J. X-LoRA: mixture of low-rank adapter experts, a flexible framework for large language models with applications in protein mechanics and design. *APL Mach. Learn.* **2** (2024).
13. Buehler, M. J. MechGPT, a language-based strategy for mechanics and materials modeling that connects knowledge across scales, disciplines and modalities. *Appl. Mech. Rev.* **76**, 021001 (2024).
14. Luu, R. K. & Buehler, M. J. BioinspiredLLM: conversational large language model for the mechanics of biological and bio-inspired materials. *Adv. Sci.* https://doi.org/10.1002/advs.202306724 (2023).
15. Cranford, S. W. & Buehler, M. J. *Biomateriomics* (Springer Netherlands, 2012).
16. Groen, N., Cranford, S., de Boer, J., Buehler, M. & Van Blitterswijk, C. *Introducing Materiomics.* (Cambridge University Press, 2011).
17. Lee, N. A., Shen, S. C. & Buehler, M. J. An automated biomateriomics platform for sustainable programmable materials discovery. *Matter* **5**, 3597–3613 (2022).
18. Arevalo, S. E. & Buehler, M. J. Learning from nature by leveraging integrative biomateriomics modeling toward adaptive and functional materials. *MRS Bull.* **48**, 1140–1153 (2023).
19. Smith, P. T., Wahl, C. B., Orbeck, J. K. & Mirkin, C. A. Megalibraries: supercharged acceleration of materials discovery. *MRS Bull.* **48**, 1172–1183 (2023).
20. Buehler, M. J. Accelerating scientific discovery with generative knowledge extraction, graph-based representation, and multimodal intelligent graph reasoning. *Mach. Learn. Sci. Technol.* **5**, 035083 (2024).
21. Hu, E. J. et al. LoRA: low-rank adaptation of large language models. *ICLR* 1, 3. Preprint at https://arxiv.org/abs/2106.09685v2 (2022).
22. Buehler, M. J. MechGPT, a language-based strategy for mechanics and materials modeling that connects knowledge across scales, disciplines and modalities. *Appl. Mech. Rev.* https://doi.org/10.1115/1.4063843 (2023).
23. Siriwardhana, S. et al. Domain adaptation of llama3-70b-instruct through continual pre-training and model merging: a comprehensive evaluation. Preprint at https://arxiv.org/abs/2406.14971 (2024).
24. Rafailov, R. et al. Direct preference optimization: your language model is secretly a reward model. *Adv. Neural Inf. Process. Syst.* **36**, 53728–53741 (2023).
25. Hong, J., Lee, N. & Thorne, J. Orpo: monolithic preference optimization without reference model. Preprint at https://arxiv.org/abs/2403.07691 (2024).
26. Christiano, P. et al. Deep reinforcement learning from human preferences. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4302–4310, Preprint at https://arxiv.org/abs/1706.03741 (2017).
27. Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. & Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of DNNs. *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Preprint at https://arxiv.org/abs/1802.10026 (2018).
28. Utans, J. Weight averaging for neural networks and local resampling schemes. In *Proc. AAAI-96 Workshop on Integrating Multiple Learned Models* (AAAI Press, 1996).
29. Goddard, C. et al. Arcee's mergekit: a toolkit for merging large language models. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Preprint at https://arxiv.org/abs/2403.13257 (2024).
30. Shoemake, K. Animating rotation with quaternion curves. *ACM SIGGRAPH Comput. Graph.* **19**, 245–254 (1985).

31. Grassia, F. S. Practical parameterization of rotations using the exponential map. *J. Graph. Tools* **3**, 29–48 (1998).
32. Hanson, A. J. Visualizing quaternions. *IEEE Comput. Graph. Appl.* **26**, 6–13 (2006).
33. Eberly, D. H. *Game Physics* (Morgan Kaufmann, 2003).
34. Belkin, M., Hsu, D., Ma, S. & Mandal, S. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc. Natl Acad. Sci. USA* **116**, 15849–15854 (2019).
35. Hansen, L. K. & Salamon, P. Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.* **12**, 993–1001 (1990).
36. Blecher, L., Cucurull, G., Scialom, T., Stojnic, R. & Ai, M. Nougat: neural optical understanding for academic documents. Preprint at https://arxiv.org/abs/2308.13418v1 (2023).
37. Liang, P. et al. Holistic evaluation of language models. Preprint at https://arxiv.org/abs/2211.09110 (2022).
38. Srivastava, A. et al. Beyond the imitation game: quantifying and extrapolating the capabilities of language models. *Trans. Mach. Learn. Res.* https://openreview.net/forum?id=uyTL5Bvosj (2023).
39. Chiang, C.-H. & Lee, H.-y. Can large language models be an alternative to human evaluations? *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Preprint at https://arxiv.org/abs/2305.01937 (Association for Computational Linguistics, 2023).
40. Zhang, Y. et al. Llmeval: a preliminary study on how to evaluate large language models. In *the proceedings of the AAAI Conference on Artificial Intelligence by the Association for the Advancement of Artificial Intelligence (AAAI)*. Vol. 38, 19615–19622 (2024).
41. Chern, S., Chern, E., Neubig, G. & Liu, P. Can large language models be trusted for evaluation? Scalable meta-evaluation of LLMs as evaluators via agent debate. Preprint at https://arxiv.org/abs/2401.16788 (2024).
42. Achiam, J. et al. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
43. Ruiz, N. et al. Dreambooth: fine tuning text-to-image diffusion models for subject-driven generation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Preprint at https://arxiv.org/abs/2208.12242 (2023).
44. Kellert, S. R., Heerwagen, J. H. & Mador, M. L. *Biophilic Design: The Theory, Science, and Practice of Bringing Buildings to Life* (Wiley, 2008).
45. Wright, F. L.*The Natural House* (Horizon Press, 1954).
46. Buehler, M. J. Cephalo: Multi-modal vision-language models for bio-inspired materials analysis and design. *Adv. Funct. Mater.* **34**, 2409531 (2024).
47. Olah, C. et al. Scaling monosemanticity. *Transform. Circ.* https://transformer-circuits.pub/2024/scaling-monosemanticity/ (2024).
48. run-llama/llama_index: LlamaIndex (formerly GPT Index) is a data framework for your LLM applications. https://github.com/run-llama/llama_index (2022).
49. Kaplan, J. et al. Scaling laws for neural language models. Preprint at https://arxiv.org/abs/2001.08361 (2020).
50. Klunk, C. L., Heethoff, M., Hammel, J. U., Gorb, S. N. & Krings, W. Mechanical and elemental characterization of ant mandibles: consequences for bite mechanics. *Interface Focus* **14**, 20230056 (2024).
51. El-Shekeil, Y. A., AL-Oqla, F. M., Refaey, H. A., Bendoukha, S. & Barhoumi, N. Investigating the mechanical performance and characteristics of nitrile butadiene rubber date palm fiber reinforced composites for sustainable bio-based materials. *J. Mater. Res. Technol.* **29**, 101–108 (2024).
52. Mbabali, H., Lubwama, M., Yiga, V. A., Were, E. & Kasedde, H. Development of rice husk and sawdust mycelium-based bio-composites: optimization of mechanical, physical and thermal properties. *J. Inst. Eng. Ser. D* **105**, 97–117 (2024).
53. Morankar, S. et al. Cholla cactus wood (cylindropuntia imbricata): hierarchical structure and micromechanical properties. *Acta Biomater.* **174**, 269–280 (2024).
54. Sauer, K. et al. Nanocrystal residual strains and density layers enhance failure resistance in the cleithrum bone of evolutionary advanced pike fish. *Acta Biomater.* **179**, 164–179 (2024).
55. Tunstall, L. et al. Zephyr: Direct distillation of lm alignment. Preprint at https://arxiv.org/abs/2310.16944 (2023).
56. Lu, W., Kaplan, D. L. & Buehler, M. J. Generative modeling, design, and analysis of spider silk protein sequences for enhanced mechanical properties. *Adv. Funct. Mater.* https://doi.org/10.1002/adfm.202311324 (2023).

Acknowledgements

This work was supported in part by Google, the MIT Generative AI Initiative, USDA (grant number 2021-69012-35978), with additional support from NIH. This material is partially based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant number 2141064.

Author contributions

M.J.B. designed the overall research, developed algorithms, and codes, and conducted the training, assessments, and analysis. M.J.B. developed and executed the distillation strategies to generate structured and adversarial data. W.L. and R.L. curated the scientific papers for the corpus of raw training data, downloaded the papers, and designed the benchmark questions. M.J.B. wrote the initial paper draft with input from W.L. and R.L., and all authors edited and finalized the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-025-01564-y.

Correspondence and requests for materials should be addressed to Markus J. Buehler.

Wei Lu[1,2], **Rachel K. Luu**[1,3] **& Markus J. Buehler**[1,2,4] ✉

[1]Laboratory for Atomistic and Molecular Mechanics (LAMM), Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [4]Center for Computational Science and Engineering, Schwarzman College of Computing, Massachusetts Institute of Technology, Cambridge, MA, USA. ✉e-mail: mbuehler@mit.edu