JRadiEvo: A Japanese Radiology Report Generation Model Enhanced by Evolutionary Optimization of Model Merging

Kaito Baba, Ryota Yagi, Junichiro Takahashi, Risa Kishikawa, Satoshi Kodera

Department of Cardiovascular Medicine
The University of Tokyo Hospital, Tokyo, Japan

baba-kaito662@g.ecc.u-tokyo.ac.jp

Abstract

With the rapid advancement of large language models (LLMs), foundational models (FMs) have seen significant advancements. Healthcare is one of the most crucial application areas for these FMs, given the significant time and effort required for physicians to analyze large volumes of patient data. Recent efforts have focused on adapting multimodal FMs to the medical domain through techniques like instruction-tuning, leading to the development of medical foundation models (MFMs). However, these approaches typically require large amounts of training data to effectively adapt models to the medical field. Moreover, most existing models are trained on English datasets, limiting their practicality in non-English-speaking regions where healthcare professionals and patients are not always fluent in English. The need for translation introduces additional costs and inefficiencies. To address these challenges, we propose a **J**apanese **Radi**ology report generation model enhanced by **Evo**lutionary optimization of model merging (JRadiEvo). This is the first attempt to extend a non-medical vision-language foundation model to the medical domain through evolutionary optimization of model merging. We successfully created a model that generates accurate Japanese reports from X-ray images using only 50 translated samples from publicly available data. This model, developed with highly efficient use of limited data, outperformed leading models from recent research trained on much larger datasets. Additionally, with only 8 billion parameters, this relatively compact foundation model can be deployed locally within hospitals, making it a practical solution for environments where APIs and other external services cannot be used due to strict privacy and security requirements.

1 Introduction

In recent years, foundational models (FMs) have seen remarkable advancements, transforming various fields by offering more sophisticated and powerful solutions [1]. A key driver of this progress has been the rise of large language models (LLMs), which have greatly expanded the capabilities of FMs, particularly in processing and generating text with high accuracy and contextual understanding. This has sparked exponential growth in research [2], leading to the development of vision-language models that integrate visual and textual data [3–5], as well as fine-tuning approaches that enhance model performance for specific tasks [6, 7].

Healthcare is one of the most critical application areas for foundational models. The need to develop models tailored to healthcare is essential, particularly because physicians often face the challenge

NeurIPS'24 Workshop on AIM-FM: Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond.

of reviewing large volumes of medical data, such as X-rays, which can be time-consuming and demanding. Advanced FMs can help alleviate this burden by enabling quicker and more efficient diagnoses, improving the overall effectiveness of healthcare delivery and patient outcomes. In response to this need, various FMs have been fine-tuned specifically for the healthcare domain, further enhancing their accuracy and effectiveness in clinical settings [7–9].

However, despite these advancements, several challenges remain. One significant issue is that most of the models developed so far, such as LLaVA-Med [7] and MedPaLM 2 [8], are predominantly in English, whereas many healthcare professionals and patients are not always proficient in English. For these models to be truly practical, there is a pressing need to expand their capabilities to non-English languages. Relying on a two-step process, where the model first generates output in English and then translates it, can introduce additional costs and complexity, making it less efficient and accessible. Additionally, publicly available datasets that can be used to train these models, such as MIMIC-CXR [10] and IU X-Ray [11], are overwhelmingly in English, with very few datasets available in other languages. Translating the large amounts of data needed for training into other languages with high quality is a costly and resource-intensive process. This scarcity of non-English datasets makes it difficult to develop models that can handle non-English languages. Furthermore, due to privacy concerns, it is challenging to collect and use patient data for model training, further complicating the creation of such datasets. Also, the use of large models through APIs, such as GPT-4 [12], is often impractical in healthcare settings because of the stringent privacy regulations that protect patient data, which limits the deployment of these models in real-world clinical environments.

To address these challenges, this paper presents a **J**apanese **Radi**ology report generation model enhanced by **Evo**lutionary optimization of model merging [13] (JRadiEvo), a first attempt to extend a multimodal vision-language model for non-English medical text generation by utilizing evolutionary optimization of model merging [13]. JRadiEvo was developed by merging a non-medical vision-language model, medical text-to-text models, and a Japanese-language text-to-text model using an evolutionary algorithm. This innovative approach enabled the efficient creation of a Japanese radiology report generation model using only a minimal amount of Japanese-language data, addressing the critical need for non-English medical text generation in a resource-constrained environment.

Below we outline our key contributions, which aim to advance the field of multimodal foundational models in healthcare:

1. **Efficient use of limited non-English medical data**: In the context of the difficulty in collecting non-English datasets, JRadiEvo demonstrates the ability to create a non-English medical report generation model by translating and utilizing only 50 cases from publicly available English datasets. This approach highlights the efficiency of the development process, demonstrating how a non-English medical report generation model can be created using extremely limited data and annotations. Additionally, it is noteworthy that not only was the dataset used after translation limited to 50 cases, but the entire dataset used to create JRadiEvo consisted of just 50 cases. This underscores the fact that JRadiEvo efficiently utilizes a very limited amount of data, demonstrating an effective approach to handling medical data under strict privacy and security constraints.

2. **Novel application of model merging in the medical vision-language model**: Traditionally, adapting models to the medical domain has relied on fine-tuning or training from scrach. To the best of our knowledge, there are no existing study of applying model merging alone to adapt a vision-language model to the medical domain. While recent research [13] has proposed using evolutionary optimization of model merging for vision-language models, this approach has been limited to natural images. To our knowledge, no prior studies have extended this technique to medical images or other domain-specific imagery beyond natural images.

3. **Lightweight model for local deployment**: JRadiEvo is an 8B parameter model, making it lightweight enough to be deployed on local hospital computing systems without the need for external APIs. This local deployment capability addresses critical privacy and security concerns, allowing hospitals to maintain control over patient data. Additionally, given the challenges of equipping facilities with expensive GPUs proportional to patient numbers, JRadiEvo's compact size and low GPU memory requirements make it practical for widespread use.

4. **Cost-efficient training process**: JRadiEvo eliminates the need for computationally expensive backpropagation during training, enabling a far more efficient learning process compared to training a new model or fine-tuning. Additionally, by leveraging model merging instead of fine-tuning, JRadiEvo avoids the common issue of catastrophic forgetting [14–16] that often occurs during fine-tuning, allowing for a more stable and efficient development process.

2 Related work

2.1 Foundation models

Medical large language models In the medical domain, several large language models (LLMs) have been developed and fine-tuned to achieve high performance. Notable models include ChatDoctor [17], DoctorGLM [18], BioGPT [19], Med-Alpaca [20], PMC-LLaMA [21], Med-Gemini [22], Med-PaLM [23], and Med-PaLM 2 [8]. These models have demonstrated impressive capabilities in understanding and generating medical text by leveraging the power of LLMs fine-tuned for healthcare-specific tasks.

Vision-language models As for multimodal vision-language models (VLMs), prominent models like Flamingo [5], Coca [24], BLIP [25], PaLI-X [26], CogVLM [27], GPT-4V [3], and LLaVA [4] have been developed to integrate visual and textual data, pushing the boundaries of what can be achieved in multimodal learning.

Medical vision-language models Recently, there has been growing interest in extending these vision-language models to the medical domain, leading to the development of models like XrayGPT [28], MedFlamingo [9], Med-PaLM M[29], LLaVA-Med [7], and CheXagent [30]. These models incorporate LLMs into vision-language frameworks specifically designed for medical applications. LLaVA-Med [7] is a vision-language model specifically designed for the medical field by using instruction-following data generated with GPT-4 [12] to perform instruction-tuning on the LLaVA [4] model. CheXagent [30] represents a significant advancement in medical vision-language models. It constructs a large-scale instruction-tuning dataset by aggregating publicly available datasets and adding new labels to existing ones, enabling accurate chest X-ray interpretation and showcasing the potential of instruction-tuned vision-language models in medical imaging.

Despite the advancements of these models, most rely on fine-tuning or training from scratch, which requires large datasets. In the medical field, where privacy and security concerns make it difficult to create large datasets, this dependency poses a significant challenge. While several publicly available English datasets can be used for training, there are very few non-English datasets, making it difficult to develop practical models for non-English-speaking regions. JRadiEvo addresses this issue by proposing an efficient method for creating a vision-language model in a non-English language using just 50 translated examples from a public dataset. Note that not only were 50 examples translated, but the entire medical image-text dataset for creating JRadiEvo consisted of only 50 cases. Furthermore, while existing models output text in English, JRadiEvo generates reports directly in Japanese, eliminating the need for translation and demonstrating the potential for practical use in non-English-speaking regions.

2.2 Model merging

Model merging is a technique that allows the strengths of multiple pre-trained models to be combined without the need for additional training. One prominent approach involves using linear or spherical linear interpolation (SLERP [31]) to merge the weights of different fine-tuned models. Another technique, known as Task Arithmetic [32], enables manipulation of features obtained through fine-tuning by creating *task vectors*, which are derived by subtracting the weights of the original pre-trained model from those of the fine-tuned model. TIES-Merging [33] takes this concept further by addressing redundant changes in task vectors and resolving conflicts in parameter signs between multiple task vectors before merging them. This method involves a three-step process: removing small, insignificant parameter changes, resolving sign conflicts between task vectors, merging the adjusted vectors. Additionally, DARE [34] proposes randomly dropping some of the changes and rescaling the remaining ones, which can be combined with techniques like Task Arithmetic [32] and TIES-Merging [33] to enhance the merging process.

Recent research [13] has also introduced the use of evolutionary algorithms to optimize parameters within TIES-Merging [33] with DARE [34]. This optimization allows for more granular merging, such as at the level of input/output embedding layers or individual transformer blocks. While previous studies have primarily focused on merging within language models, this research extends the applicability of model merging to vision-language models, demonstrating its effectiveness in this multimodal context.

Most prior studies, except for recent work [13], have focused primarily on language models, without extending their methods to multimodal applications. Although recent work [13] introduced evolutionary algorithm-based optimization and extended model merging to vision-language models, it was limited to natural images and did not extend to domain-specific images beyond natural imagery, such as medical images. Our work is the first to extend evolutionary model merging to the medical domain, specifically for chest X-ray images, demonstrating its effectiveness in this highly specialized context.

3 JRadiEvo

JRadiEvo efficiently adapted a vision-language model (VLM) to the non-English medical domain through the evolutionary optimization of TIES-Merging [33] combined with DARE [34].

3.1 Problem setting

3.1.1 Vision-language models

VLM is designed to generate a text response y given an image x_I and accompanying text x_T. A typical VLM utilizing a large language model (LLM) is composed of three main components: a vision encoder $\mathcal{M}_V$ that extracts features from the image, a projector $\mathcal{M}_P$ that transforms these image features into the latent space of the LLM, and an LLM component $\mathcal{M}_L$ that generates the output text. The formulation of this process can be expressed as:

$$y = \mathcal{M}_L(\mathcal{M}_P(\mathcal{M}_V(x_I)), x_T). \tag{1}$$

In this setup, the LLM component $\mathcal{M}_L$ is a pre-trained model that has already acquired strong language capabilities, such as Llama 3 [35]. To adapt it for vision-related tasks, the projector $\mathcal{M}_P$, and optionally the LLM component $\mathcal{M}_L$ are trained or fine-tuned.

3.1.2 Model merging

Let $\theta_{\text{init}} \in \mathbb{R}^d$ represent the trainable parameters of the pre-trained LLM, where d is the parameter dimension. Given a set of K tasks $\{t_1, t_2, \cdots, t_K\}$, the LLM is fine-tuned on these tasks, resulting in a set of fine-tuned parameter vectors $\left\{\theta_{\text{ft}}^{t_1}, \theta_{\text{ft}}^{t_2}, \cdots, \theta_{\text{ft}}^{t_K}\right\}$. Here, each task t_i corresponds to a specific domain or task for fine-tuning the pre-trained LLM, such as vision tasks, medical applications, or adaptation to the Japanese language.

As demonstrated in previous research [32], the task vector for a task t is defined as the difference between the fine-tuned weights for task t and the original pre-trained weights, i.e., the task vector $\tau_t \in \mathbb{R}^d$ is given by:

$$\tau_t = \theta_{\text{ft}}^t - \theta_{\text{init}}.$$

This task vector corresponds to the capabilities acquired through fine-tuning on task t. By manipulating these task vectors and merging it with the original weights θ_{init}, we can merge the capabilities of multiple fine-tuned models without additional training. The details of the merging process we adopted are described in Section 3.2.

3.2 Model merging for JRadiEvo

In JRadiEvo, following the approach of previous work [13], we optimized TIES-Merging [33] combined with DARE [34] using an evolutionary algorithm. For the VLM, we also adhered to the strategy used in earlier research [13] by focusing on the parameters of the LLM component $\mathcal{M}_L$ during the merging process, i.e., $\theta_{\text{ft}}^{t_1}$ represents the parameters of the LLM component $\mathcal{M}_L$ of a VLM fine-tuned on vision-to-text data. Meanwhile, $\theta_{\text{ft}}^{t_2}, \theta_{\text{ft}}^{t_3}, \cdots, \theta_{\text{ft}}^{t_K}$ are the parameters of the LLM

Table 1: The statistics of the MIMIC-CXR dataset [10], showing the number of images, reports, and patients in each split.

	Train	Valid	Test
Image	368,960	2,991	5,159
Report	222,758	1,808	3,269
Patient	64,586	500	293

fine-tuned on text-to-text data, covering tasks $t_2, t_3, \cdots, t_K$ related to medical knowledge or the Japanese language.

The resulting parameters $\left\{\theta_{\text{ft}}^{t_1}, \theta_{\text{ft}}^{t_2}, \cdots, \theta_{\text{ft}}^{t_K}\right\}$ were then merged using the corresponding task vectors $\{\tau_{t_1}, \tau_{t_2}, \cdots, \tau_{t_K}\}$ as follows:

1. **DARE**: Following the approach outlined in previous work [34], with $\alpha \in \mathbb{R}$ as the drop rate, the following operation was performed:

$$m^t \sim \text{Bernoulli}(\alpha), \quad \tilde{\tau}_{\text{DARE}}^t = (1 - m^t) \odot \tau_{\text{ft}}^t, \quad \tau_{\text{DARE}}^t = \tilde{\tau}_{\text{DARE}}^t / (1 - \alpha),$$

   where $\odot$ denotes element-wise multiplication. This operation randomly drops some of the changes in the task vector τ_t and rescales the remaining ones.

2. **TIES-Merging**: Similar to the previous work [33], we removed small, trivial changes in the task vectors, resolved sign conflicts between them, and merged:
   (a) **Trim**: To remove insignificant changes, for each task t, we created a task vector $\tilde{\tau}_t$ by setting all parameters of a task vector τ_{DARE}^t to zero except for those absolute values in the top k_t percent.
   (b) **Elect**: To resolve sign conflicts, for each parameter $p \in \{1, 2, \cdots, d\}$, we calculated the sign with greater total movement as $\gamma_m^p = \text{sgn}\left(\sum_{t=1}^K \tilde{\tau}_t^p\right)$.
   (c) **Merge**: Finally, for each parameter p, we computed the weighted sum of only the parameters from task vectors whose signs matched the aggregated elected sign. Specifically, the merged task vector τ_{merged} is given by $\tau_{\text{merged}}^p = \frac{1}{|\mathcal{A}_p|} \sum_{t \in \mathcal{A}_p} c_t \tilde{\tau}_t^p$, where $\mathcal{A}_p = \{t \in [n] | \, \text{sgn}(\tilde{\tau}_t^p) = \gamma_m^p\}$, and $c_t \in \mathbb{R}$ is a weight assigned to each task vector. The merged task vector τ_{merged} is then scaled by a scaling parameter $\lambda \in \mathbb{R}$ and added to the initial parameters θ_{init} to obtain the final parameters: $\theta_{\text{final}} = \theta_{\text{init}} + \lambda \tau_{\text{merged}}$.

3. **Evolutionary optimization**: Following the previous work [13], we leveraged an evolutionary algorithm to optimize the parameters of the step above. Specifically, we optimized DARE drop rates α, TIES-Merging saved rates $\{k_{t_1}, k_{t_2}, \cdots, k_{t_K}\}$, weight $\{c_{t_1}, c_{t_2}, \cdots, c_{t_K}\}$, and scaling parameters λ. We treated the entire model as a single layer for the merging process, identical to the approach used in previous research [13]. We iteratively calculated θ_{final} using steps 1 and 2, with θ_{final} serving as $\mathcal{M}_L$ in Eq. (1). The parameters were suggested by the evolutionary algorithm to maximize the ROUGE-L [36] score between the generated text $\hat{y}$ from equation Eq. (1) and the reference text y. This process was repeated, with the evolutionary algorithm continuously refining the parameters to improve the score.

Finally, the text was generated using Eq. (1) with the θ_{final} obtained and optimized through this process.

4 Experiments

4.1 Experimental setup

Datasets For our experiments, we used the MIMIC-CXR [10], a publicly available dataset that consists of 377,110 chest X-ray (CXR) images and 227,835 corresponding English-language radiology reports. The images are provided in both DICOM and JPEG formats [37]. The dataset is officially split into training, validation, and test sets, containing 368,960, 2,991, and 5,159 images, respectively. The details about the dataset are presented in Table 1.

Following the previous work [38, 39], samples without corresponding reports were excluded from the dataset. Additionally, according to previous research [40], for cases where multiple images are associated with a single report, only the first image was used. From the resulting official training set, which included metadata indicating AP (anteroposterior) and PA (posteroanterior) views, we randomly selected 50 samples from both views to create the dataset used in our study. These selected 50 samples were translated into Japanese using GPT-3.5 [41][1], and the translations were then reviewed and and revised by a human to ensure accuracy.

For the test dataset, we extracted samples from the official test set. As with the training data, only those with corresponding reports were selected, the first image was used for reports associated with multiple images, and we focused exclusively on AP and PA views. These selected samples were translated into Japanese using GPT-3.5 [41][1]. Note that the training samples were drawn from the official training set and the test samples from the official test set, ensuring no data leakage occurred between the training and test phases.

Evaluation metrics Following the previous studies[39, 40, 38, 42], we evaluated the generated Japanese radiology reports using BLEU [43], ROUGE-L [36], and METEOR [44] scores. These metrics are commonly used in machine translation and text generation tasks and provide a comprehensive evaluation of the quality of the generated text. The generated Japanese reports and the translated reference texts were tokenized using MeCab [45], a Japanese-specific tokenizer, before calculating the evaluation metrics.

Source models To efficiently create a Japanese VLM capable of understanding medical content through evolutionary model merging, we merged a non-medical VLM fine-tuned on vision tasks, two text-to-text LLMs fine-tuned on medical datasets, and another text-to-text LLM fine-tuned on Japanese datasets. Specifically, we used `Bunny-v1_1-Llama-3-8B-V`[2] [46] as the VLM, `MMed-Llama-3-8B-EnIns`[3] [47] and `OpenBioLLM-Llama3-8B`[4] [48] as the medical models, and `Llama-3-Swallow-8B-Instruct-v0.1`[5] [49] as the Japanese language model. All of them are the fine-tunes of the Llama 3 [35].

Evolutionary optimization For the evolutionary algorithm described in Section 3.2, we used CMA-ES [50] implemented in Optuna [51], following the previous study [13]. As hyperparameters for the CMA-ES algorithm, all parameters were initialized to 0.5, with a sigma value of 1/6, and a population size of $4 + \lfloor 3 \ln(n) \rfloor$, where n is the number of parameters. The algorithm was run for 600 iterations, and the best parameters were selected based on the ROUGE-L [36] score. We provided the prompt x_T, "You are a skilled radiologist. Please examine this X-ray image and write a report. Pay attention to any abnormalities in the lungs, heart, or bones. Answer in Japanese," written in Japanese.

4.2 Comparison with leading models from previous study and instruction-tuning approaches

To evaluate the effectiveness of our evolutionary model merging approach, we compared the results with those obtained by LoRA [52] instruction-tuning the same vision-language model (VLM). Additionally, we compared our results with the performance of leading models from recent research.

4.2.1 Experimental conditions

As the base VLM, we used `Bunny-v1_1-Llama-3-8B-V`[6] [46], the same model used in the merging process in JRadiEvo, and the fine-tuning. During this process, the weights of the image encoder $\mathcal{M}_V$ were kept fixed, and we prepared two models: one where only the LLM component $\mathcal{M}_L$ was tuned, and another where both the LLM $\mathcal{M}_L$ and projector $\mathcal{M}_P$ were tuned. Both models were trained using LoRA [52]. For the LoRA setup, the hyperparameters were set with a decomposition rank r of 8 and a scaling factor α of 16. The learning rate was was set to 2×10^{-4} and decayed with a cosine annealing [53]. The model was trained for one epoch with a batch size of 8 and gradient

[1] We accessed through Azure OpenAI service.

[2] `https://huggingface.co/BAAI/Bunny-v1_1-Llama-3-8B-V`

[3] `https://huggingface.co/Henrychur/MMed-Llama-3-8B-EnIns`

[4] `https://huggingface.co/aaditya/Llama3-OpenBioLLM-8B`

[5] `https://huggingface.co/tokyotech-llm/Llama-3-Swallow-8B-Instruct-v0.1`

[6] `https://huggingface.co/BAAI/Bunny-v1_1-Llama-3-8B-V`

Table 2: Comparison of JRadiEvo with leading models from recent research and instruction-tuned approaches. **Bold** and underlined scores are the best and worst in each metric, respectively.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
JRadiEvo (ours)	0.376	0.301	0.241	0.179	**0.212**	**0.191**
LLM instruction-tuned	0.342	0.241	0.185	0.133	0.143	0.142
LLM + projector instruction-tuned	0.345	0.245	0.189	0.136	0.147	0.146
CheXagent [30]	0.221	0.181	0.149	0.119	0.199	0.159
GPT-4o [54]	**0.432**	**0.333**	**0.257**	**0.184**	0.176	0.188

accumulation steps set to 2. The training was conducted using a single NVIDIA A100 GPU with 80GB memory.

As for the dataset, we followed the same procedure outlined in Section 4.1, using the training data from MIMIC-CXR [10]. Specifically, samples without corresponding reports were excluded, only the first image was used for reports with multiple images, and we focused on AP and PA view images. From this dataset, 2,000 samples were randomly selected and translated into Japanese using GPT-3.5 [41][1]. This translated dataset was then used for instruction-tuning.

For comparison with previous studies, we used CheXagent [30], a recent leading model specifically designed for chest X-ray images. CheXagent [30], proposed in 2024, is one of the most recent advancements in the field. It was trained on a large instruction-tuning dataset, which was created by aggregating publicly available data and adding new labels to existing datasets. Since CheXagent is trained in English, reports were first output in English using CheXagent and then translated into Japanese using GPT-4o [54][1]. This translated Japanese reports were used for comparison.

As another comparison, we used GPT-4o [54][1], a high-performance VLM that can output directly in Japanese.

4.2.2 Results and discussion

The results are shown in Table 2. Also, an example comparison of generated text from JRadiEvo and the ground truth on the test data is shown in Table 3.

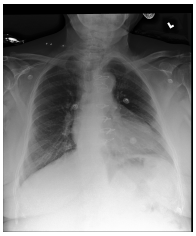
Effectiveness of JRadiEvo in generating radiology reports We can see from this table that our model, JRadiEvo, achieved the highest scores in both ROUGE-L and METEOR metrics. Given that ROUGE-L is considered the most aligned with human judgment in evaluating generated text, as shown in previous research [36], this underscores the effectiveness of our application of evolutionary model merging to medical text generation.

Efficient use of limited datasets When compared to the latest model, CheXagent [30], JRadiEvo outperformed it across all evaluation metrics. Despite CheXagent being trained on a vast instruction-tuning dataset, JRadiEvo, with only 50 training samples, significantly surpassed it in generating X-ray reports. This demonstrates JRadiEvo's ability to effectively utilize a extremely limited dataset to create a powerful medical foundation model.

Practicality in non-English-speaking regions Additionally, while CheXagent [30] produces reports in English, our model eliminates the need for an additional translation step by directly generating reports in Japanese. This is particularly important in non-English-speaking medical environments, where neither doctors nor patients may be fluent in English. Requiring translation every time is time-consuming and costly. Thus, JRadiEvo demonstrates potential for practical use in non-English-speaking regions.

Parameter efficiency and local deployment In comparison with GPT-4o [54], while it scored higher on BLEU, JRadiEvo outperformed it on METEOR and ROUGE-L, which is more closely aligned with human evaluation. This shows that the performance is comparable to, or even surpasses, that of GPT-4o. Considering that JRadiEvo is a lightweight model with only 8 billion parameters, whereas GPT-4o is significantly larger, this highlights JRadiEvo's impressive parameter efficiency. Furthermore, unlike GPT-4o, which requires API access, JRadiEvo's modest size allows it to be

Table 3: Example comparison of generated text from JRadiEvo and the ground truth on the test data. (For reference, the English translation by GPT-4o is shown in parentheses.)

Example 1		
	Generated	レントゲン写真を読影した結果、肺野に多数の小さな 白斑が見られます。これは肺炎の可能性があります。 心臓の形状は正常です。骨には異常は見られません。 (The X-ray examination shows multiple small white spots in the lung fields, which may indicate pneumonia. The shape of the heart is normal, and no abnormalities are observed in the bones.)
	Ground truth	肺容量は低下している。左肺上葉と比較し、右上葉により強く瘢痕性、慢性の線維化が見られる。左肺中部に線状影が見られ、左下葉にも軽度見られる。胸水はない。心臓、縦郭の輪郭は正常である。 (The lung volume is reduced. Compared to the left upper lobe, the right upper lobe shows more pronounced scarring and chronic fibrosis. Linear opacities are observed in the mid-left lung, and mild findings are also present in the left lower lobe. There is no pleural effusion. The contours of the heart and mediastinum are normal.)
Example 2		
	Generated	レントゲン写真を読影した結果、心臓の左室が大きく、右室が小さいことがわかりました。肺野は均一に空気を含み、骨は正常です。心臓の左室の異常は心不全の可能性を示唆しています。 (The X-ray examination revealed that the left ventricle of the heart is enlarged, while the right ventricle is small. The lung fields are uniformly aerated, and the bones are normal. The abnormality in the left ventricle suggests the possibility of heart failure.)
	Ground truth	胸骨正中切開のワイヤーと縦郭のクリップが見られる。肺野は清明で、浸潤影や胸水、気胸はない。心臓や縦郭陰影は正常。骨病変はない。右横隔膜下にフリーエアはない。 (Wires from a median sternotomy and clips in the mediastinum are visible. The lung fields are clear, with no infiltrates, pleural effusion, or pneumothorax. The cardiac and mediastinal silhouettes are normal. There are no bone lesions. No free air is seen under the right diaphragm.)
Example 3		
	Generated	レントゲン写真を読影した結果、左肺下葉に不規則な影が見られます。心臓の形状は正常です。骨には異常は見られません。左肺下葉の影は肺炎や肺癌などの可能性があります。詳細な検査が必要です。 (The X-ray examination shows an irregular opacity in the left lower lobe. The shape of the heart is normal, and no abnormalities are observed in the bones. The opacity in the left lower lobe may suggest possibilities such as pneumonia or lung cancer. Further detailed examination is required.)
	Ground truth	ポータブル撮影画像。左肺下部の病変が拡大しており、胸水と浸潤影を疑う。肺野に散在する結節影は悪性腫瘍の転移を疑う。心陰影は不明瞭で心拡大の評価は困難。骨病変は指摘できず。 (Portable imaging shows an enlarged lesion in the lower left lung, raising suspicion of pleural effusion and infiltrates. Scattered nodules in the lung fields suggest possible metastasis of a malignant tumor. The cardiac silhouette is unclear, making it difficult to assess for cardiomegaly. No bone lesions are noted.)

deployed locally in hospitals. This makes JRadiEvo a more practical option for use in privacy- and security-sensitive environments in hospitals.

Superiority over instruction-tuning Comparing JRadiEvo with the two models that instruction-tuned the same VLM shows that JRadiEvo outperforms them across all metrics. This highlights the effectiveness of adopting evolutionary model merging over traditional instruction-tuning approaches. Moreover, increasing the amount of data used for instruction-tuning in an attempt to further enhance the model can lead to catastrophic forgetting [14–16]. In our experiments, instruction-tuning with 2,000 data points yielded good results, but when we increased the dataset to 10,000 data points, catastrophic forgetting occurred, severely impairing the language functionality and rendering the model unusable. As previous research [55] has indicated, instruction tuning with LoRA [52] in Japanese is not effective for relatively small models, a finding that our results also confirm. This suggests that similar challenges may arise in other non-English languages as well. In contrast, JRadiEvo leveraged evolutionary model merging to successfully adapt the source VLM to the medical domain without experiencing catastrophic forgetting. This highlights the potential of model merging as a viable approach for domain adaptation in non-English languages, especially for smaller models.

4.3 Analysis of merged LLM contributions

To investigate the contributions of the merged LLMs, we compared the density and weight parameters after optimization, following the approach outlined in previous research [13]. Specifically, we

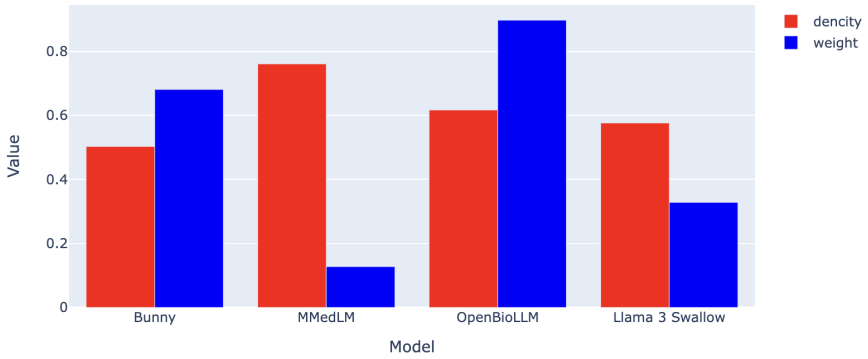


Figure 1: Evolved configurations for each models.

analyzed the retained percentage $\{k_{t_1}, k_{t_2}, \cdots, k_{t_K}\}$ and the merging weight $\{c_{t_1}, c_{t_2}, \cdots, c_{t_K}\}$ described in Section 3.2. The results of this comparison are presented in Figure 1.

As shown in Figure 1, the weight for OpenBioLLM [48] is significantly higher, and its density is also relatively high. This suggests that the medical knowledge embedded in OpenBioLLM was crucial for adapting the non-medical VLM to the medical domain. In contrast, while MMedLM [47] has a high density, its weight is much lower, indicating that its contribution was less significant. This suggests that, in this setup, OpenBioLLM's medical knowledge was primarily utilized, potentially rendering MMedLM less influential in the model merging process.

Additionally, when examining the density and weight of Llama 3 Swallow [49], the LLM enhanced for Japanese language proficiency, we see that it was utilized to some extent, though not as heavily as OpenBioLLM. This suggests that the original VLM and LLMs had a some capacity for handling Japanese, albeit imperfectly, and the lack of medical knowledge was a more significant limitation. However, unlike MMedLM, which saw a dramatic reduction in weight, Llama 3 Swallow's contribution was not negligible, indicating that its role was still necessary. In fact, when asked directly in Japanese, the other VLM and LLMs could generate responses that, while somewhat awkward and unnatural from a native speaker's perspective, were still intelligible.

5 Conclusion

In this study, we proposed a Japanese Radiology report generation model enhanced by Evolutionary optimization of model merging (JRadiEvo), marking the first attempt to extend a multimodal vision-language model for non-English medical text generation using evolutionary model merging. Despite utilizing only 50 translated samples from publicly available data, JRadiEvo demonstrated superior performance compared to leading models from recent studies trained on much larger datasets. This highlights the effectiveness of our approach in efficiently leveraging limited data to create a powerful and practical medical foundation model.

While JRadiEvo has shown promising results in evaluation metrics, human judgment by medical experts or further refinement may be needed to make it suitable for clinical use. Future work includes efforts to close this gap to ensure the model's reliability in real-world medical settings.

6 Acknowledgments

This work was supported by Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" Grant Number JPJ012425.

References

[1] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al., "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[2] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[3] Open AI, "GPT-4V(ision) system card," 2023.

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds. 2023, vol. 36, pp. 34892–34916, Curran Associates, Inc.

[5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan, "Flamingo: a visual language model for few-shot learning," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds. 2022, vol. 35, pp. 23716–23736, Curran Associates, Inc.

[6] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao, "Instruction tuning with GPT-4," *arXiv preprint arXiv:2304.03277*, 2023.

[7] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao, "LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds. 2023, vol. 36, pp. 28541–28564, Curran Associates, Inc.

[8] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan, "Towards expert-level medical question answering with large language models," *arXiv preprint arXiv:2305.09617*, 2023.

[9] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar, "Med-Flamingo: a multimodal medical few-shot learner," in *Proceedings of the 3rd Machine Learning for Health Symposium*, Stefan Hegselmann, Antonio Parziale, Divya Shanmugam, Shengpu Tang, Mercy Nyamewaa Asiedu, Serina Chang, Tom Hartvigsen, and Harvineet Singh, Eds. 10 Dec 2023, vol. 225 of *Proceedings of Machine Learning Research*, pp. 353–367, PMLR.

[10] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, no. 1, pp. 317, 2019.

[11] Dina Demner-Fushman, Marc Kohli, Marc Rosenman, Sonya Shooshan, Laritza Rodriguez, Sameer Antani, George Thoma, and Clement Mcdonald, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association : JAMIA*, vol. 23, 07 2015.

[12] OpenAI, "GPT-4 Technical Report," *arXiv preprint arXiv:2303.08774*, 2024.

[13] Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha, "Evolutionary optimization of model merging recipes," *arXiv preprint arXiv:2403.13187*, 2024.

[14] Michael McCloskey and Neal J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," vol. 24 of *Psychology of Learning and Motivation*, pp. 109–165. Academic Press, 1989.

[15] Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan, "Understanding catastrophic forgetting in language models via implicit inference," in *The Twelfth International Conference on Learning Representations*, 2024.

[16] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang, "An empirical study of catastrophic forgetting in large language models during continual fine-tuning," *arXiv preprint arXiv:2308.08747*, 2024.

[17] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang, "Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge," *Cureus*, vol. 15, 06 2023.

[18] Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen, "DoctorGLM: Fine-tuning your chinese doctor is not a herculean task," *ArXiv*, vol. abs/2304.01097, 2023.

[19] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu, "BioGPT: generative pre-trained transformer for biomedical text generation and mining," *Briefings in Bioinformatics*, vol. 23, no. 6, pp. bbac409, 09 2022.

[20] Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem, "MedAlpaca – an open-source collection of medical conversational ai models and training data," *arXiv preprint arXiv:2304.08247*, 2023.

[21] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie, "Pmc-llama: Towards building open-source language models for medicine," *arXiv preprint arXiv:2304.14454*, 2023.

[22] Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan, "Capabilities of gemini models in medicine," *arXiv preprint arXiv:2404.18416*, 2024.

[23] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[24] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu, "Coca: Contrastive captioners are image-text foundation models," *Transactions on Machine Learning Research*, 2022.

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proceedings of the 39th International Conference on Machine Learning*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, Eds. 17–23 Jul 2022, vol. 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900, PMLR.

[26] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut, "PaLI-X: On scaling up a multilingual vision and language model," *arXiv preprint arXiv:2305.18565*, 2023.

[27] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang, "Cogvlm: Visual expert for pretrained language models," 2023.

[28] Omkar Chakradhar Thawakar, Abdelrahman M. Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Khan, "XrayGPT: Chest radiographs summarization using large medical vision-language models," in *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa, Kirk Roberts, and Junichi Tsujii, Eds., Bangkok, Thailand, Aug. 2024, pp. 440–448, Association for Computational Linguistics.

[29] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan, "Towards generalist biomedical ai," *arXiv preprint arXiv:2307.14334*, 2023.

[30] Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S Chaudhari, and Curtis Langlotz, "Chexagent: Towards a foundation model for chest x-ray interpretation," in *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

[31] Tom White, "Sampling generative networks," *arXiv preprint arXiv:1609.04468*, 2016.

[32] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi, "Editing models with task arithmetic," in *The Eleventh International Conference on Learning Representations*, 2023.

[33] Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal, "TIES-merging: Resolving interference when merging models," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[34] Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li, "Language models are super mario: Absorbing abilities from homologous models as a free lunch," in *International Conference on Machine Learning*. PMLR, 2024.

[35] AI@Meta, "Llama 3 model card," 2024.

[36] Chin-Yew Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Barcelona, Spain, July 2004, pp. 74–81, Association for Computational Linguistics.

[37] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng, "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.

[38] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. 2018, vol. 31, Curran Associates, Inc.

[39] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan, "Generating radiology reports via memory-driven transformer," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, Eds., Online, Nov. 2020, pp. 1439–1449, Association for Computational Linguistics.

[40] Aaron Nicolson, Jason Dowling, and Bevan Koopman, "Improving chest X-ray report generation by leveraging warm starting," *Artificial Intelligence in Medicine*, vol. 144, pp. 102633, 2023.

[41] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 1877–1901, Curran Associates, Inc.

[42] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert, "Interactive and explainable region-guided radiology report generation," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7433–7442, 2023.

[43] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, USA, 2002, ACL '02, p. 311–318, Association for Computational Linguistics.

[44] Michael Denkowski and Alon Lavie, "Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar F. Zaidan, Eds., Edinburgh, Scotland, July 2011, pp. 85–91, Association for Computational Linguistics.

[45] Taku Kudo, "MeCab : Yet another part-of-speech and morphological analyzer," 2006.

[46] Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao, "Efficient multimodal learning from data-centric perspective," *arXiv preprint arXiv:2402.11530*, 2024.

[47] Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie, "Towards building multilingual language model for medicine," 2024.

[48] Malaikannan Sankarasubbu Ankit Pal, "OpenBioLLMs: Advancing open-source large language models for healthcare and life sciences," `https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B`, 2024.

[49] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki, "Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities," in *First Conference on Language Modeling*, 2024.

[50] Nikolaus Hansen, *The CMA Evolution Strategy: A Comparing Review*, pp. 75–102, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[51] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2623–2631.

[52] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[53] Ilya Loshchilov and Frank Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations*, 2017.

[54] Open AI, "GPT-4o system card," 2024.

[55] Hao Wang, Akifumi Nakamachi, and Toshinori Sato, "Lora tuning for large-scale japanese foundational models," in *Proceedings of the 29th Annual Meeting of the Association for Natural Language Processing*, Tokyo, Japan, March 2023, The Association for Natural Language Processing.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the contributions made in the paper, including the adaptation of a vision-language model to the non-English medical domain through the evolutionary optimization of model mearging.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of this work at the end of the conclusion, noting the need for human judgment and further refinement for clinical use.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides sufficient information to reproduce the experimental results, including the model merging process, the evolutionary optimization process, and the LoRA instruction-tuning process.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The paper does not provide open access to the code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides sufficient information on the training and test details, including the data splits, hyperparameters, optimizer, and how they were chosen.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not report error bars or statistical significance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides sufficient information on the compute resources needed to reproduce the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both potential positive and negative societal impacts of the work performed, including the potential for improved medical diagnostics and the need for further refinement for clinical use.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The paper describes safeguards, including the need for human judgment and further refinement for clinical use.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly credits the creators of existing assets and mentions the license and terms of use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer:[NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.