nature medicine

Article

https://doi.org/10.1038/s41591-024-03185-2

A generalist vision–language foundation model for diverse biomedical tasks

Received: 29 January 2024

Accepted: 10 July 2024

Published online: 07 August 2024

Check for updates

Kai Zhang[1], Rong Zhou[1], Eashan Adhikarla[1], Zhiling Yan[1], Yixin Liu[1], Jun Yu[1], Zhengliang Liu[2], Xun Chen[3], Brian D. Davison[1], Hui Ren[4], Jing Huang[5,6], Chen Chen[7], Yuyin Zhou[8], Sunyang Fu[9], Wei Liu[10], Tianming Liu[2], Xiang Li[4] ✉, Yong Chen[5,11,12,13], Lifang He[1] ✉, James Zou[14,15], Quanzheng Li[4], Hongfang Liu[9] & Lichao Sun[1] ✉

Traditional biomedical artificial intelligence (AI) models, designed for specific tasks or modalities, often exhibit limited flexibility in real-world deployment and struggle to utilize holistic information. Generalist AI holds the potential to address these limitations due to its versatility in interpreting different data types and generating tailored outputs for diverse needs. However, existing biomedical generalist AI solutions are typically heavyweight and closed source to researchers, practitioners and patients. Here, we describe BiomedGPT, the first open-source and lightweight vision–language foundation model, designed as a generalist capable of performing various biomedical tasks. BiomedGPT achieved state-of-the-art results in 16 out of 25 experiments while maintaining a computing-friendly model scale. We also conducted human evaluations to assess the capabilities of BiomedGPT in radiology visual question answering, report generation and summarization. BiomedGPT exhibits robust prediction ability with a low error rate of 3.8% in question answering, satisfactory performance with an error rate of 8.3% in writing complex radiology reports, and competitive summarization ability with a nearly equivalent preference score to human experts. Our method demonstrates that effective training with diverse data can lead to more practical biomedical AI for improving diagnosis and workflow efficiency.

AI techniques, especially transformer-based foundation models, have demonstrated their power in solving a wide range of biomedical tasks, including radiology interpretation, clinical-information summarization and precise disease diagnostics[1]. However, most of today's biomedical models act as specialist systems, tailored to specific tasks and modalities[2]. Such specialization comes with substantial challenges in model deployment, especially with the growing interest in using AI for precision medicine and patient-centered care, which require the integration and analysis of diverse data types and patient-specific details[3,4]. Furthermore, the hyper-specialization of AI in narrow disciplines often fails to provide the comprehensive insights necessary to assist doctors in real-world settings, where the flow of information can be slow and sporadic[2,5]. A generalist biomedical AI has the potential to overcome these limitations by using versatile models that can be applied to different tasks and are robust enough to handle the intricacies of medical data effectively[2,6].

The emergence of general-purpose foundation models[7,8] offers a prototype for the development of biomedical generalist AI. These advanced models serialize diverse datasets, regardless of their modalities, tasks or domains, into a uniform sequence of tokens, which are then processed using a transformer neural network[9]. Unlike large language models[10,11], which are primarily designed for processing textual data, generalist models can handle both textual and visual information simultaneously. This capability is pivotal for complex biomedical

A full list of affiliations appears at the end of the paper. ✉e-mail: xli60@mgh.harvard.edu; lih319@lehigh.edu; lis221@lehigh.edu

applications, in which the integration of diverse data types—such as clinical text and radiographic imaging—is crucial for accurate analysis and decision-making. Furthermore, generalist models exhibit impressive multitasking capabilities, greatly simplifying the deployment and management of AI systems by reducing the need to maintain numerous narrowly focused specialist models.

In this paper, we introduce BiomedGPT, a prototype for a generalist vision–language foundation model designed to perform diverse biomedical tasks across modalities using natural-language instructions (Fig. 1). Unlike multimodal biomedical AI systems that are specialized for a single task[12], focused solely on one discipline[13] or not publicly accessible[6], BiomedGPT is trained with cross-disciplinary data and evaluated on a wide range of tasks. BiomedGPT is fully transparent, open-source and lightweight (for example, it is 3,088 times smaller than the commercial generalist biomedical AI model Med-PaLM M, which has 562 billion parameters[6]), thereby facilitating broader implementation. To empower the generalist capabilities of BiomedGPT, we curated a large-scale pretraining corpus comprising 592,567 images, approximately 183 million text sentences, 46,408 object–label pairs and 271,804 image–text pairs (Fig. 2c,d). Furthermore, to enhance its ability to follow instructions, we developed a variant called Instruct-BiomedGPT with specifically curated instruction-tuning data (Supplementary Fig. 1).

To our knowledge, BiomedGPT is the first fully transparent generalist medical AI model that has been comprehensively evaluated on publicly accessible datasets and by medical professionals. This study first highlights the transfer-learning capabilities of BiomedGPT, demonstrating how the model uses knowledge from pretraining to specialize effectively across 25 datasets through fine-tuning (Extended Data Tables 1 and 2 and Supplementary Table 7). We used recognized metrics from the literature to benchmark our model against state-of-the-art (SOTA) results. Additionally, BiomedGPT is a zero-shot learner that can answer multimodal medical questions without further training for adaptation, and its performance is comparable to that of leading AI systems. Furthermore, doctors evaluated BiomedGPT in tasks such as visual question answering (VQA), report generation and summarization within the radiology domain, and it demonstrated satisfactory performance. Although our results highlight BiomedGPT's potential in medical applications, they also indicate that substantial enhancements are required to make it usable in the clinic. Critical evaluations for BiomedGPT are particularly needed in the areas of safety, equity and bias. Our findings underscore the challenges that must be addressed before these models can be deployed effectively in clinical settings. We outline these limitations and suggest directions for future research.

Results

Pretraining using large and diverse datasets

BiomedGPT uses pretraining techniques including masked modeling and supervised learning, aiming to establish robust and general data representations by learning from extensive datasets across diverse tasks (Extended Data Table 3). To maximize the generalization of BiomedGPT, we sourced the pretraining data from 14 freely available datasets, ensuring the diversity of modalities (Figs. 1a and 2c,d and Extended Data Fig. 1a). In addition, to investigate how BiomedGPT performs across scales, we specifically introduced three versions of the model: BiomedGPT-S, BiomedGPT-M and BiomedGPT-B, which correspond to small, medium and base sizes, respectively (Fig. 2a and Extended Data Figs. 2 and 3).

Fine-tuning for downstream tasks

Multitasking is fundamental to a generalist AI. Following previous biomedical research[14–16] and aiming for sufficiently effective performance, we primarily fine-tuned our model to adapt to various biomedical tasks (Fig. 1b,c). Our selection of downstream tasks stemmed from their potential real-world applications: medical-image classification can aid in disease diagnostics and lesion recognition; text understanding and summarization can streamline clinic operations, such as easing doctors' note-writing burden. Furthermore, image captioning and VQA lay the groundwork for future healthcare chatbots, addressing challenges in which common language might be ambiguous but medical terminology is too complex for most people to understand. The complete statistics of downstream datasets used in this article are shown in Extended Data Figure 1b.

BiomedGPT is lightweight but competitive in multimodal tasks

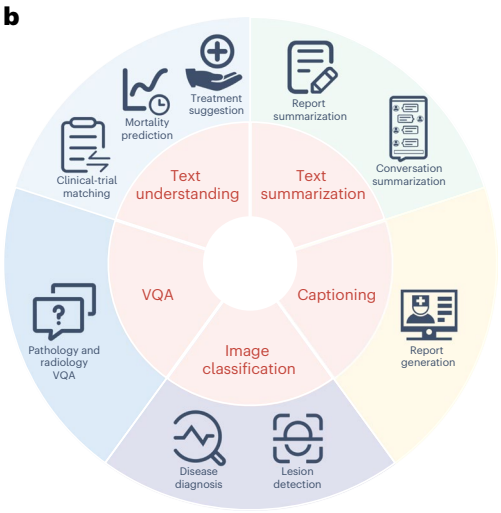
We fine-tuned BiomedGPT on two primary multimodal tasks, VQA and image captioning, each using three downstream datasets. The VQA datasets included radiology data covering five anatomies (VQA-RAD[17] and Semantically-Labeled Knowledge-Enhanced Dataset (SLAKE)[18]), in addition to pathology data that captures both anatomical and tissue-specific details (PathVQA[19]). For captioning, we incorporated chest X-ray (CXR) datasets (IU X-ray[20] and Medical Information Mart for Intensive Care III-CXR (MIMIC-CXR)[21]) as well as clinical photographs from Peir Gross[22]. For comparison, we benchmarked BiomedGPT against leading models for each dataset[15,23–25].

We evaluated our model's VQA performance by comparing generated answers with the ground truths. The overall accuracy of our BiomedGPT model is detailed in Extended Data Table 1. Notably, BiomedGPT achieved an 86.1% overall accuracy on the SLAKE dataset, surpassing the previous state-of-the-art (SOTA) performance of 85.4%, set by BiomedCLIP[15]. Additionally, we dissected the accuracy of both 'closed ended' and 'open ended' question–answer pairs (Fig. 3a). Our model recorded promising closed-ended accuracies: 88.0% on PathVQA, up by 1.0% compared with the performance of the current SOTA model[25]. On the SLAKE dataset, BiomedGPT-B achieved an 89.9% closed-ended accuracy, down by 1.1% compared with the M2I2 model's performance[23]. In open-ended scenarios, our model excelled with an 84.3% accuracy, surpassing M2I2's 74.7%. However, for the VQA-RAD and PathVQA datasets, BiomedGPT's performance on open-ended queries was less competitive, recording accuracies of 60.9% and 28.0%, respectively.

In addition, we compared BiomedGPT-B with Med-PaLM M (12 billion parameters) using the weighted F_1 score, as reported in the paper[6]. Other metrics could not be calculated owing to the closed-source nature of Med-PaLM M. Remarkably, despite its much smaller size, BiomedGPT-B achieved impressive results (Fig. 2b). On the VQA-RAD and SLAKE datasets, BiomedGPT-B attained scores of 73.2% and 85.2%, respectively, which represent a substantial increase of 22.5% on VQA-RAD and a slight improvement of 0.02% on SLAKE. Additionally, on the PathVQA dataset, BiomedGPT-B had a weighted F_1 score of 56.9%, only 0.4% lower than Med-PaLM M, while utilizing a model with 98.5% fewer parameters.

To evaluate the model's image-captioning ability (Fig. 3b), we meticulously assessed the quality of machine-generated text using three metrics: recall-oriented understudy for gisting evaluation-longest common subsequence (ROUGE-L)[26], metric for evaluation of translation with explicit ordering (METEOR)[27] and consensus-based image description evaluation (CIDEr)[28]. We compared the performance of BiomedGPT to that of established models[13,29–33]. These evaluation metrics are useful for assessing the similarity and consensus between the generated text and the reference text written by medical experts. They have also shown some alignment with ratings given by physicians[34]. Consequently, models that score higher on these natural-language processing (NLP) metrics can be selected as candidates for further human evaluation[35]. On the Peir Gross dataset, our BiomedGPT model surpassed the existing SOTA benchmark[36], demonstrating improvements of 8.1 percentage points in ROUGE-L and 0.5 points in METEOR, and a substantial gain of 89.8 points in the CIDEr metric. Conversely, on the IU X-ray dataset, BiomedGPT achieved a leading CIDEr score

https://doi.org/10.1038/s41591-024-03185-2



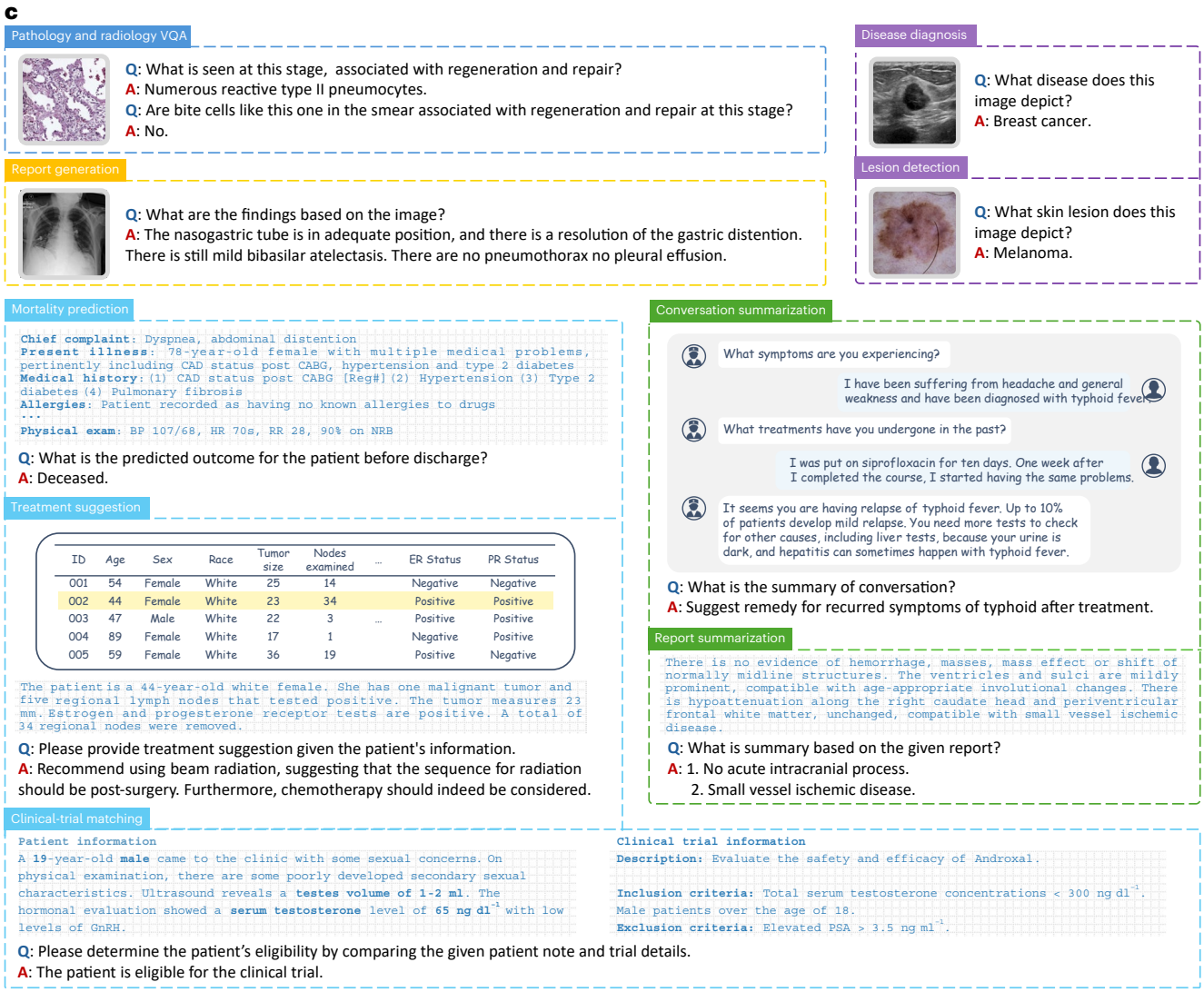


Fig. 1 | BiomedGPT can process diverse modalities and perform versatile tasks. a, BiomedGPT focuses primarily on visual and textual inputs, but can also process tabular data through serialization. CT, computed tomography; EHR, electronic health records; EKG, electrocardiogram; MRI, magnetic resonance imaging. **b**, Examples of the supported downstream visual-language tasks of BiomedGPT demonstrate its versatility. Additional tasks can be incorporated to meet further clinical needs through lightweight, task-specific fine-tuning. **c**, Examples of clinically relevant use-cases for BiomedGPT include tasks in which the input consists of both image and text or only text; the model responds to queries (Q) by generating responses (A). Thanks to its unified framework design and comprehensive pretraining on biomedical data, BiomedGPT is highly adaptable and can be applied to a variety of downstream tasks. BP, blood pressure; CABG, coronary artery bypass graft surgery; CAD, coronary artery disease; ER, estrogen receptor; GnRH, gonadotropin-releasing hormone; HR, heart rate; NRB, non-rebreather mask; PR, progesterone receptor; RR, respiratory rate; Reg#, de-identified 'Medical Record Number'.

https://doi.org/10.1038/s41591-024-03185-2

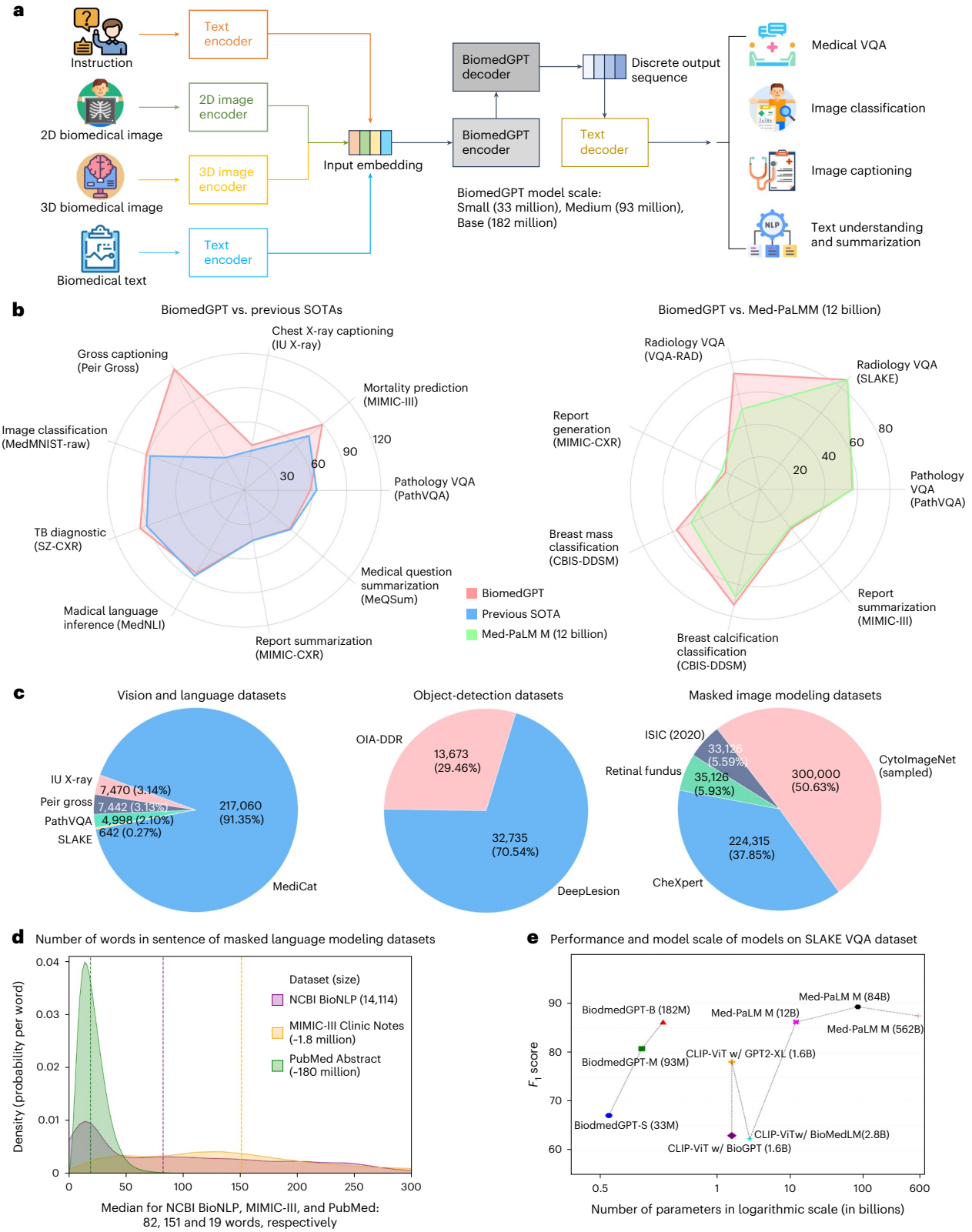


Fig. 2 | An overview of BiomedGPT: workflow, performance and pretraining datasets. a, Illustration of how BiomedGPT handles multimodal inputs and performs diverse downstream tasks. The expected form of output for each task is determined by feeding the specific instruction to the model. 2D, two-dimensional. **b**, Comparative performance analysis contrasting the achievements of BiomedGPT with prior SOTA results and Med-PaLM M (12 billion parameters). The evaluation metrics include accuracy for image classification, medical language inference and VQA (benchmarked against SOTA results); CIDEr for image captioning; ROUGE-L for text summarization; weighted F_1 scores for VQA (in comparison with Med-PaLM M); and F_1-macro for breast mass and calcification classification (also in comparison with Med-PaLM M). **c**, Distribution of pretraining datasets including image captioning and VQA as vision and language datasets, object-detection datasets and image-only datasets for masked image modeling. **d**, Density plot of the number of words per sentence in the text-only pretraining datasets. **e**, A comparison of scale-related performance. BiomedGPT exhibits superior performance on the SLAKE VQA dataset, although it has considerably fewer parameters than its counterparts. B, billion; M, million.

https://doi.org/10.1038/s41591-024-03185-2

a

Medical VQA performance

Model	Parameters	VQA-RAD accuracy		SLAKE accuracy		PathVQA accuracy	
		Closed-ended	Open-ended	Closed-ended	Open-ended	Closed-ended	Open-ended
BiomedGPT-S (ours)	33M (0.2×)	57.8 (23.5↓)	13.4 (47.5↓)	73.3 (16.6↓)	66.5 (17.8↓)	84.2 (3.8↓)	10.7 (17.3↓)
BiomedGPT-M (ours)	93M (0.5×)	79.8 (1.5↓)	53.6 (7.3↓)	86.8 (3.1↓)	78.3 (6.0↓)	85.7 (2.3↓)	12.5 (15.5↓)
M2I2	252M (1.4×)	81.6 (0.3↑)	61.8 (0.9↑)	91.1 (0.2↑)	74.7 (9.6↓)	88.0	36.3 (8.3↑)
BiomedCLIP	422M (2.3×)	79.8 (1.5↓)	67.6 (6.7↑)	89.7 (0.2↓)	82.5 (1.8↓)	-	-
CLIP-ViT with GPT2-XL	1.6B (8.8×)	-	-	82.1 (7.8↓)	84.3	87.0 (1.0↓)	40.0 (12.0↑)
MedVInT-TD	7.0B (38.5×)	86.8 (5.5↑)	73.7 (12.8↑)	86.3 (3.6↓)	84.5 (0.2↑)	-	-
BiomedGPT-B (ours)	182M	81.3	60.9	89.9	84.3	88.0	28.0

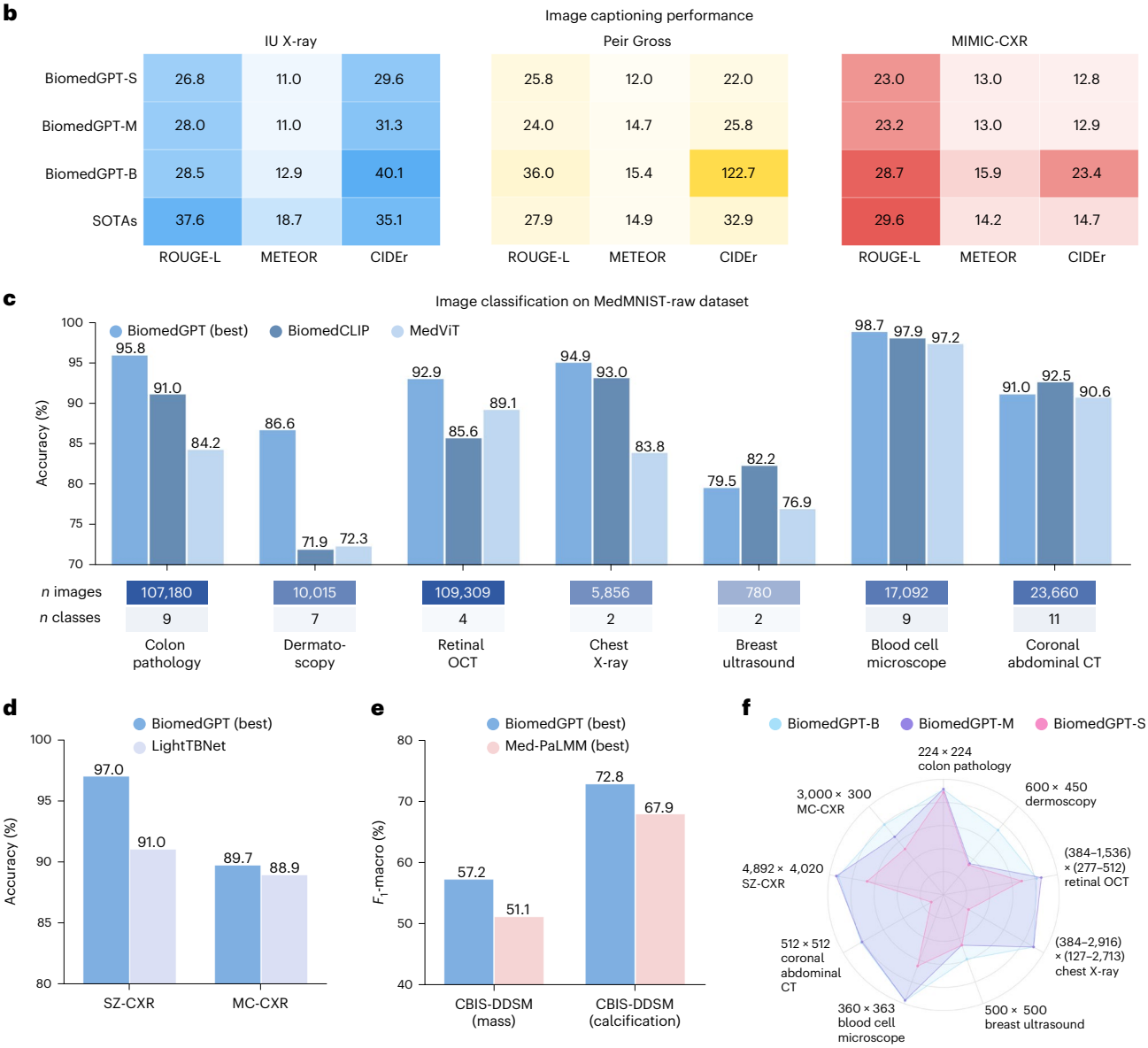


Fig. 3 | BiomedGPT performs fine-tuning for vision–language and medical-image-classification downstream tasks. a, Medical VQA performance of BiomedGPT and the leading models, in terms of closed-ended and open-ended accuracies. The information in parentheses indicates the performance change compared to BiomedGPT-B. × denotes the multiple of the parameter size of other models relative to that of our model. ↓ denotes the performance decrease compared to our model. ↑ denotes the performance increase compared to our model. For example, 0.5↓ means that the corresponding model has 0.5 lower accuracy than BiomedGPT-B. **b**, Image-captioning performance of BiomedGPT and SOTA platforms on IU X-ray, Peir Gross and MIMIC-CXR data. The evaluation metrics are ROUGE-L, METEOR and CIDEr. **c**, Evaluation of image classification on the MedMNIST-Raw dataset for each domain type. **d**, Image-classification performance with accuracy across two super-resolution image datasets. **e**, Image-classification performance as assessed by the F_1-macro on the CBIS-DDSM dataset. **f**, Accuracies across nine datasets with different resolutions (shown on the graph, in pixels) vary with model scale. In general, larger models tend to perform better.

of 40.1, marking a 5.0-point improvement over the SOTA model[31]. On the MIMIC-CXR dataset, in terms of METEOR, our model recorded a score of 15.9%, surpassing the previous leading result[30].

BiomedGPT enables accurate medical-image classification

For the medical-image-classification task, we curated a biomedical image dataset, named MedMNIST-Raw, encompassing seven modalities following ref. 37: (1) colon pathology with nine tissue types; (2) dermoscopy images of seven typical pigmented skin lesions; (3) breast ultrasound (normal, benign and malignant); (4) retinal optical coherence tomography (OCT) categorized into four types of retinal diseases; (5) CXR images for binary-class classification of pneumonia against normal; (6) blood cell microscope showcasing eight kinds of normal cells; and (7) abdominal computed tomography (CT) with 11 body organs across the coronal view. Additionally, we tested the model on two super-resolution pulmonary disease datasets, with a specific focus on pulmonary tuberculosis (TB), which has a limited number of samples: (8) the Montgomery County CXR set (MC-CXR), with dimensions of either 4,020 × 4,892 or 4,892 × 4,020 pixels; and (9) the Shenzhen CXR set (SZ-CXR), with approximate dimensions of 3,000 × 3,000 pixels. To be consistent with prior works, we used accuracy for evaluation. As shown in Figure 3c–e, BiomedGPT outperformed previous SOTA systems on seven of the nine biomedicalimage-classification datasets after five-epoch fine-tuning.

Notably, on the SZ-CXR and MC-CXR datasets[38] (binary classification), BiomedGPT had accuracies of 97.0% and 89.7%, reflecting improvements of 6.0% and 0.8%, respectively, over the previously leading model, LightTBNet[39] (Fig. 3d). For MedMNIST-Raw, we selected two top-performing approaches on biomedical imaging analysis, MedViT (Large)[40] and BiomedCLIP[15], as benchmarks for comparison. For BiomedCLIP, we added a decision layer and fine-tuned the entire model. BiomedGPT achieved 5 out of 7 best accuracies on MedMNIST-Raw (Fig. 3c): for example, on the dermoscopy dataset, BiomedGPT surpassed the two baseline models by more than 14%. On average, BiomedGPT achieved performance improvements of 6.1% and 3.3% over MedViT and BiomedCLIP, respectively.

BiomedGPT exhibits performance enhancements as its scale increases (Fig. 3f). Specifically, on the MC-CXR dataset, the small model had an accuracy of 75.9%. By contrast, the medium model had a score of 82.8%, which is 6.9% higher than its smaller counterpart's performance. The base model continued this upward trajectory, with a score of 89.7%, surpassing the medium model by 6.9%. However, we also observed performance saturation on several datasets, such as SZ-CXR. We also tested the extreme situation in which the images were resized to a very small scale and found that performance saturation became much more pronounced (Supplementary Table 1).

Additionally, we benchmarked BiomedGPT against Med-PaLM M on the Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) dataset[41] for both three-class lesion-level mass classification and calcification classification. Using the macro-averaged F_1 score (F_1-macro) as the evaluation metric, consistent with how Med-PaLM M was evaluated, we found that BiomedGPT-B outperforms all versions of Med-PaLM M, spanning 12 billion, 84 billion and 584 billion parameters (Fig. 3e and Extended Data Fig. 4a). These findings underscore the impressive efficiency and efficacy of BiomedGPT, even relative to models with larger scales.

BiomedGPT understands and summarizes clinical text

We assessed BiomedGPT's proficiency in understanding and condensing complex medical narratives that hold potential for addressing real-world clinical needs: (1) medical natural-language inference, using the MedNLI dataset[42], which tests the model's comprehension in deducing hypotheses from provided premises; (2) treatment suggestions for radiation therapy and chemotherapy based on the Surveillance, Epidemiology, and End Results (SEER) dataset[43]; (3) in-hospital mortality prediction on the basis of admission notes; and (4) clinical-trial matching that identifies lists of candidate clinical trials suitable for individuals. Moreover, we explored BiomedGPT's performance in medical-text summarization, which was applied to datasets of doctor–patient dialogues (MedQSum[44] and HealthCareMagic[45]) as well as radiology reports (MIMIC-CXR[21] and MIMIC-III[46]).

While evaluating the MedNLI dataset for three-class classification (entailment, contradiction or neutral), we used accuracy as our evaluation metric, consistent with prior research (Fig. 4e). Notably, when compared with the SOTA performance of SciFive-Large[16] at 86.6% accuracy, BiomedGPT-B, which has merely a quarter of SciFive-Large's parameter count, exhibited a decline in accuracy of only 2.8%.

For the treatment-suggestion task, we adopted the preprocessing steps as described in prior work[47]. An example output is: 'Recommend using beam radiation, suggesting that the sequence for radiation should be post-surgery. Furthermore, chemotherapy should indeed be considered.' To evaluate the effectiveness of three variants in treatment suggestions, we used a tenfold cross-validation method and compared current open-source SOTA methods, including BioGPT[14] and LLaVA-Med[12] (Fig. 4a), which have 347 million and 7 billion parameters, respectively—approximately 11 and 212 times larger, respectively, than BiomedGPT-S. BiomedGPT-B achieved a mean accuracy of 50.0% ± 5.3%, outperforming BioGPT and LLaVA-Med, which had accuracies of 45.9% ± 4.8% and 41.5% ± 7.1%, respectively. Considering the complexity involved with six types of radiation therapy, seven radiation sequences and two types of chemotherapy[47], which together imply a random-guess accuracy of 1.2%, both BiomedGPTs and the baseline models have much higher accuracies than this baseline.

For the clinical-trial matching task, we collected a dataset from Text Retrieval Conference (TREC) 2022[48], categorized into three groups: eligible, irrelevant and ineligible. We randomly chose 80% of the data from each group as the training set and the remaining 20% as the test set, and reported the average results across 10 repetitions. Again, all three versions of BiomedGPT outperformed the baselines (Fig. 4b). In particular, BiomedGPT-B achieved a mean accuracy of 85.2% ± 1.5%, substantially outperforming BioGPT and LLaVA-Med, which had accuracies of 42.0 % ± 1.8% and 48.7% ± 2.4%, respectively.

To assess BiomedGPT's performance in predicting in-hospital mortality, we used admission notes extracted from the MIMIC-III database, following ref. 49, with the official test set. Figure 4c presents the prediction-accuracy results for five models, demonstrating that all three versions of BiomedGPT outperformed BioGPT and LLaVA-Med. Notably, BiomedGPT-B achieved an accuracy improvement of more than 15% compared with these two baselines.

We used the ROUGE-L metric to assess BiomedGPT-B's text-summarization performance across four benchmark datasets (Fig. 4d). BiomedGPT-B demonstrated its ability to summarize doctor–patient dialogues on the MedQSum and HealthCareMagic datasets, achieving ROUGE-L scores of 52.3% and 42%, respectively. Leading models[32], with 400 million parameters (at least twice as large as BiomedGPT-B), recorded ROUGE-L scores of 53.2% and 44.7%, BiomedGPT-B showed only minor performance drops of 0.9% and 2.7%. Additionally, in summarizing radiology reports, and specifically in generating impressions from radiologists' findings, BiomedGPT-B achieved a ROUGE-L score of 44.4% on the MIMIC-CXR dataset. This result is closely aligned with the performance of the SOTA model, trailing by a mere 0.1% from the top score of 44.5%[33]. In the MIMIC-III dataset, BiomedGPT-B's performance stood out with a ROUGE-L score of 30.7%, surpassing Med-PaLM M (12 billion parameter), which scored 29.5%.

BiomedGPT can perform zero-shot prediction on new data

We focused on evaluating the zero-shot capabilities of BiomedGPT in VQA, highlighting its ability to answer biomedical questions in a freeform manner at scale, without requiring retraining. This contrasts sharply with earlier biomedical AI models, such as bidirectional encoder

https://doi.org/10.1038/s41591-024-03185-2

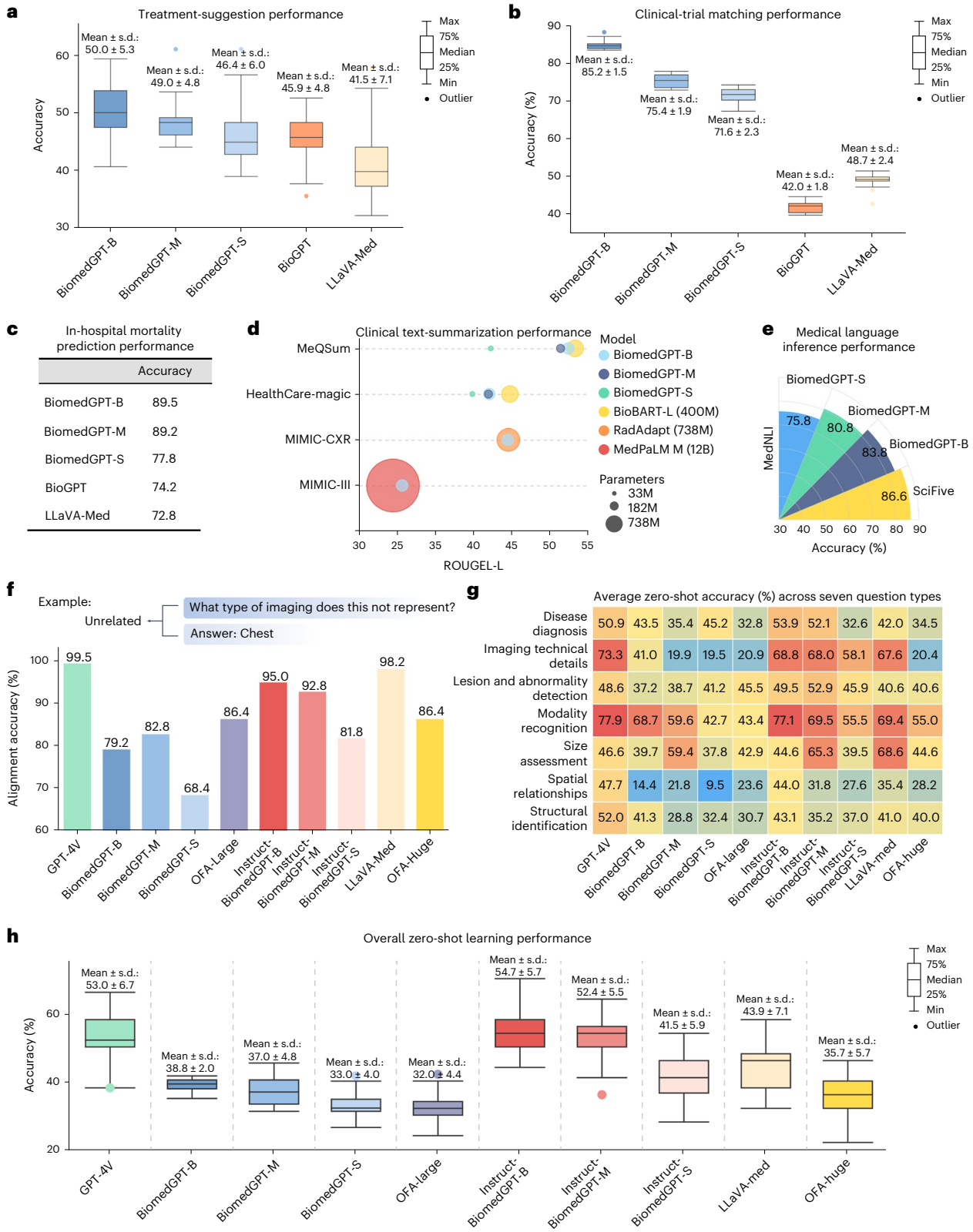


Fig. 4 | BiomedGPT performs few-epoch transfer learning for clinical-text understanding and summarization and generates a response through zero-shot transfer learning. a, Evaluation of models for the treatment-suggestion task in terms of accuracy using tenfold cross-validation ($n = 4{,}680$ data samples). **b**, Comparison of performance, assessed using accuracy, on the patient–trial matching dataset, derived from the TREC 2022 dataset, using tenfold cross-validation ($n = 7079$ data samples). **c**, Accuracy across three BiomedGPT variants and two SOTA models, BioGPT and LLaVA-Med, for in-hospital mortality prediction. **d**, ROUGE-L scores across four text-summarization datasets, relative to model scale. **e**, Medical language inference performance on the MedNLI dataset. **f**, Comparison of zero-shot question-alignment accuracy among Instruct-BiomedGPTs (base, medium, small), BiomedGPTs, OFAs (large, huge), LLaVA-Med and GPT-4V. An example illustrating a mismatch between the generated answer and the question is shown. **g**, Average zero-shot accuracy across seven question types on the VQA-RAD dataset. **h**, Overall zero-shot learning performance on the VQA-RAD dataset over 50 repeated samplings ($n = 39$ data samples).

representations from transformers (BERT)-based or vision transformer (ViT)-based models[40], which are incapable of zero-shot prediction, or contrast language–image pretraining (CLIP)-based models[15], which require predefined answers (Extended Data Fig. 5a). Unlike these models, BiomedGPT can generate answers by simply processing the input data, offering more flexible and dynamic AI-driven solutions for biomedical inquiries. In addition to medical VQA, BiomedGPT showcased zero-shot capabilities in disease diagnosis and X-ray report generation, matching the performance of Med-PaLM M and LLaVA-Med (Extended Data Fig. 5b,c).

We used the VQA-RAD dataset[18] (which was absent from the pretraining data) for evaluation, through 50 random samplings. Our evaluation of BiomedGPT's performance centered on two key metrics: (1) the accuracy of the model in providing correct answers, and (2) its ability to understand the questions and respond in a contextually relevant way, measured as alignment accuracy. We noted low alignment accuracy, indicating poor question comprehension, by our pretrained models (Fig. 4f). To address this, we developed Instruct-BiomedGPT which was fine-tuned using instruction-tuning data (Supplementary Fig. 1). We assessed this model against current SOTA models, including GPT-4V[50], LLaVA-Med (7B)[12], OFA-Huge (930 million parameters) and OFA-Large (470 million parameters)[51] in a zero-shot setting, analyzing various question types (Extended Data Table 4). Specifically, Instruct-BiomedGPT-B achieved a zero-shot accuracy of 54.7% ± 5.7%, surpassing GPT-4V's 53.0% ± 6.7% (Fig. 4h). Despite this improvement in understanding medical questions, neither model reached clinically acceptable performance. For example, the current top-performing medical vision–language model, LLaVA-Med, achieved accuracies of only 42.0% and 40.6% in disease diagnosis and lesion detection, respectively (Fig. 4g). Although Instruct-BiomedGPT-B showed a more than 10% improvement over LLaVA-Med, accuracies remained under 60%. These results highlight the complexity of diagnosis and the need for ongoing fine-tuning in the development of visual-language biomedical AI.

Regarding alignment accuracy, GPT-4V and LLaVA-Med outperformed the other models (Fig. 4f); specifically, they achieved impressive scores of 99.5% ± 1.1% and 98.2% ± 2.0%, respectively, likely owing to the advanced large language models on which they are built[10,11]. The marked improvement in alignment accuracy between Instruct-BiomedGPT and the pretrained BiomedGPT exemplifies the effectiveness of instruction tuning in enhancing the model's capability to follow instructions accurately. For instance, BiomedGPT-B achieved a mean alignment accuracy of 79.2%, but Instruct-BiomedGPT-B reached 95%.

Human evaluation of BiomedGPT for radiology tasks

To evaluate the clinical applicability and deployment challenges of BiomedGPT, we conducted a series of analyses through radiologist evaluations of the model's generated responses to a wide range of tasks, including VQA, report generation and report summarization in radiology. Examples of human evaluation on these three tasks in terms of response factuality, omissions and severity of the errors are shown in Figure 5a. The detailed evaluation procedure and performance analysis are as follows.

Radiology VQA. To clinically evaluate the correctness of BiomedGPT's responses, we randomly selected 52 question–answer samples from 16 images in the official test set of MIMIC-Diff-VQA[52] over 6 categories (Supplementary Table 2): abnormality, presence, location, type, view and severity level. For a fair comparison, we collected the answers generated by BiomedGPT, LLaVA-Med after fine-tuning and GPT-4V (zero-shot). The generated answers were presented to a seasoned radiologist at Massachusetts General Hospital for scoring (Fig. 5b,c). The answers were categorized as correct, partially correct, incorrect or unrelated, and were assigned scores of 2, 1, 0 and −1, respectively. Additionally, the original radiology reports were provided to the radiologist to serve as a reference, potentially facilitating a more precise evaluation.

BiomedGPT achieved an average score of 1.75 across all 52 samples, accumulating a total score of 91. In comparison, GPT-4V and LLaVA-Med attained average scores of 1.17 and 1.4, resulting in total scores of 61 and 73, respectively. BiomedGPT demonstrated superior performance in four out of five question categories. In addition, despite the radiologist identifying some errors in the sampled gold labels from MIMIC-Diff-VQA, we conducted a comparison using an exact match score based on these labels across the test set with non-difference questions. In this evaluation, BiomedGPT-B showed the best performance (Supplementary Table 3).

Radiology report generation. This task's complexity arises from the need for long-form outputs that provide detailed descriptions of various aspects, such as the presence, location and severity of abnormalities. In this study, we randomly selected 30 sample image–report pairs from the MIMIC-CXR dataset[21]. We then applied BiomedGPT-B and BiomedGPT-M to generate the 'findings' section of the radiology report based on the input CXR image. The radiologist assessed the quality of the generated text by addressing several aspects. First, they identified any disagreements with the generated report, such as incorrect finding locations, incorrect severity levels, references to views not present or mentions of prior studies that do not exist. Second, the radiologist determined whether the errors in the generated report are critical, with the options being critical, noncritical or N/A if more information is needed. Third, they pinpointed any omissions in the generated text. Finally, the radiologist judged whether the omissions are clinically critical.

In the evaluation, we focused on finding-level metrics, in which the generated text would be split into individual findings. For instance, the report 'PA and lateral views of the chest provided. Cardiomegaly is again noted with mild pulmonary edema. No large effusion or pneumothorax.' consists of three findings. To clearly demonstrate the quality of the generated findings, we quantified the error rates and omission rates (Fig. 5d). In the analysis of 192 generated findings, BiomedGPT-B achieved a rate of 'critical error' of 8.3%, whereas BiomedGPT-M exhibited a rate of 11.0% (excluding one case that required additional information for a comprehensive impact assessment). These rates are comparable to the human observer variabilities on the MIMIC-CXR, which has an error rate of approximately 6%[53]. We also reported the rate of 'harmless error'; BiomedGPT-B and BiomedGPT-M achieved 5.2% and 11.5%, respectively. Our observations included an analysis of 254 findings from the reference report to calculate the omission rates. The total omission rates for BiomedGPT-B and BiomedGPT-M were 23.3% and 23.5%, respectively. Because not all findings described in the reference are clinically necessary, our analysis primarily focused on critical omissions; BiomedGPT-B and BiomedGPT-M had similar rates, of 7.0% and 6.9%, respectively.

Radiology report summarization. We evaluated 100 summaries generated by BiomedGPT-B based on findings from MIMIC-CXR data[21], alongside the 'Impression' sections of corresponding reference reports. Our evaluation focused on completeness, correctness and potential medically adverse effects due to any omissions or incorrect interpretations (Fig. 5a). Completeness is rated from 1 (very incomplete) to 5 (very complete), with 3 representing a borderline (neutral) encapsulation. Accuracy is assessed by how well the content reflects the clinical implications for the patient, rated from 1 (very incorrect) to 5 (very correct). The potential for medically adverse effects from errors is classified as 'no harm', 'mild' or 'severe', on the basis of their clinical impact. Finally, we compared which summary, generated or referenced, better encapsulated all clinically relevant information, providing a comprehensive comparison of AI-generated summaries with traditional radiology reports in terms of relevance, accuracy and safety.

https://doi.org/10.1038/s41591-024-03185-2

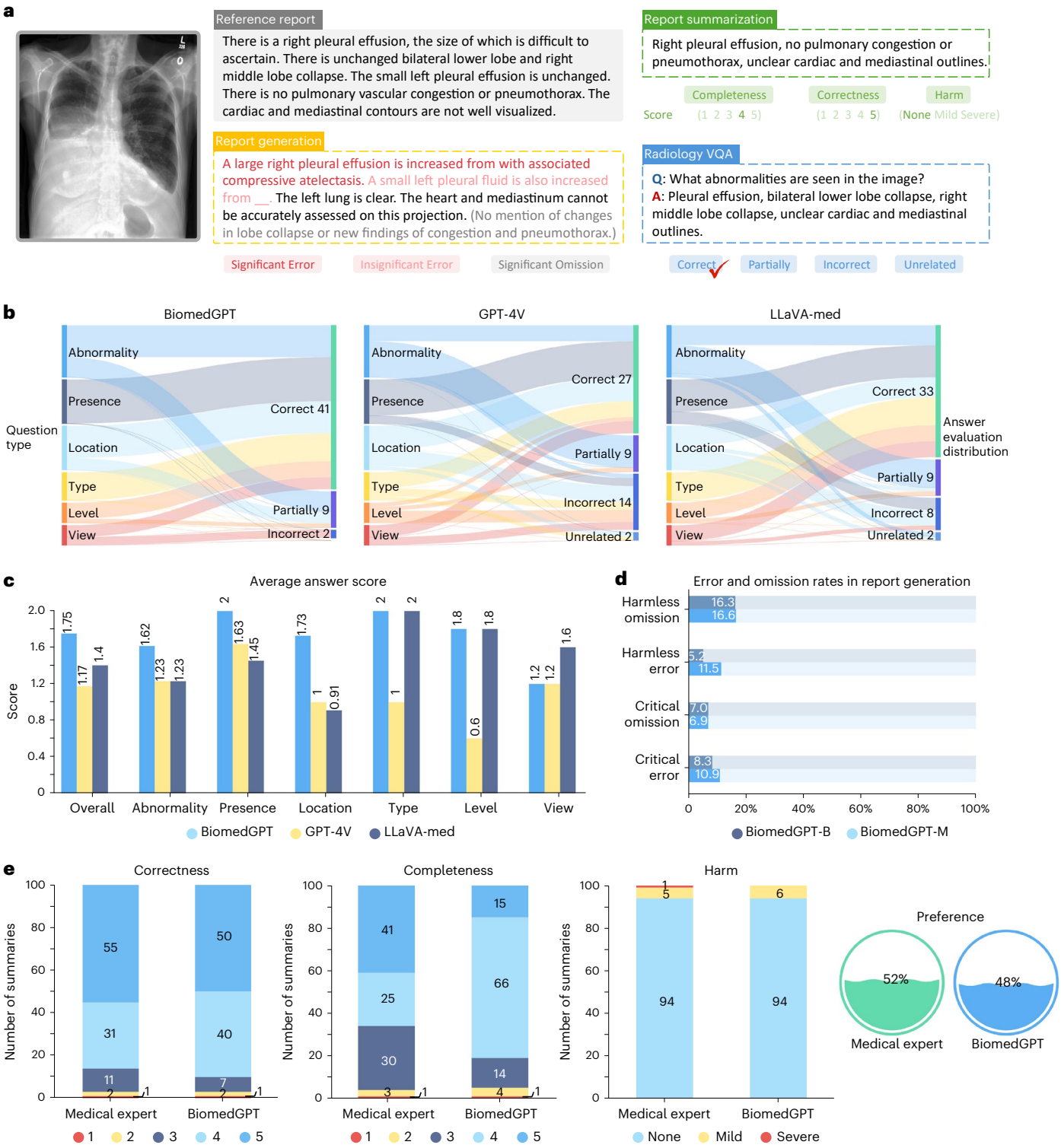


Fig. 5 | Human evaluation of the VQA, text-summarization and captioning tasks. a, Examples of human evaluation for three tasks in terms of response factuality, omissions and severity of the errors. In the given X-ray image, L indicates the left side of the patient's body; the 'O' is not a letter but the imaging of a foreign object either inside or outside the subject's body. **b**, Comparison of performance between three models across six question categories for radiology VQA. **c**, Average answer score for radiology VQA. **d**, Error and omission rates of BiomedGPT-B and BiomedGPT-M in the generated radiology report. **e**, Human evaluation of report summarization considers three attributes: completeness, correctness and potential harm, with the radiologist's preference. Specifically, in all comparison pairs (reference summary from the medical expert and the BiomedGPT-generated summary, the radiologist evaluator prefer the reference summary in 52% of cases. For the remaining 48% of the cases, the evaluator think the BiomedGPT-generated summary is better.).

BiomedGPT-generated summaries generally exhibit higher completeness (Fig. 5e), achieving average completeness (score > 3) in 81.0% of cases, 15.0% higher than the reference summaries. Additionally, only 5% of BiomedGPT-generated summaries are considered incomplete (score < 3), compared with 4% for the reference summaries. Despite these findings, the average completeness score for BiomedGPT is

https://doi.org/10.1038/s41591-024-03185-2

b

	BloodMNIST	DermaMNIST	SLAKE-MRI	SLAKE-CT
RadGPT-S	93.2 (4.5↓)	68.1 (10.5↓)	25.9 (42.5↓)	42.0 (38.5↓)
RadGPT-M	96.5 (1.2↓)	67.8 (10.8↓)	43.4 (25.0↓)	57.2 (23.3↓)
RadGPT-B	96.8 (0.9↓)	70.5 (8.1↓)	66.2 (2.2↓)	65.3 (15.2↓)
BiomedGPT-B (ours)	97.7	78.6	68.4	80.5

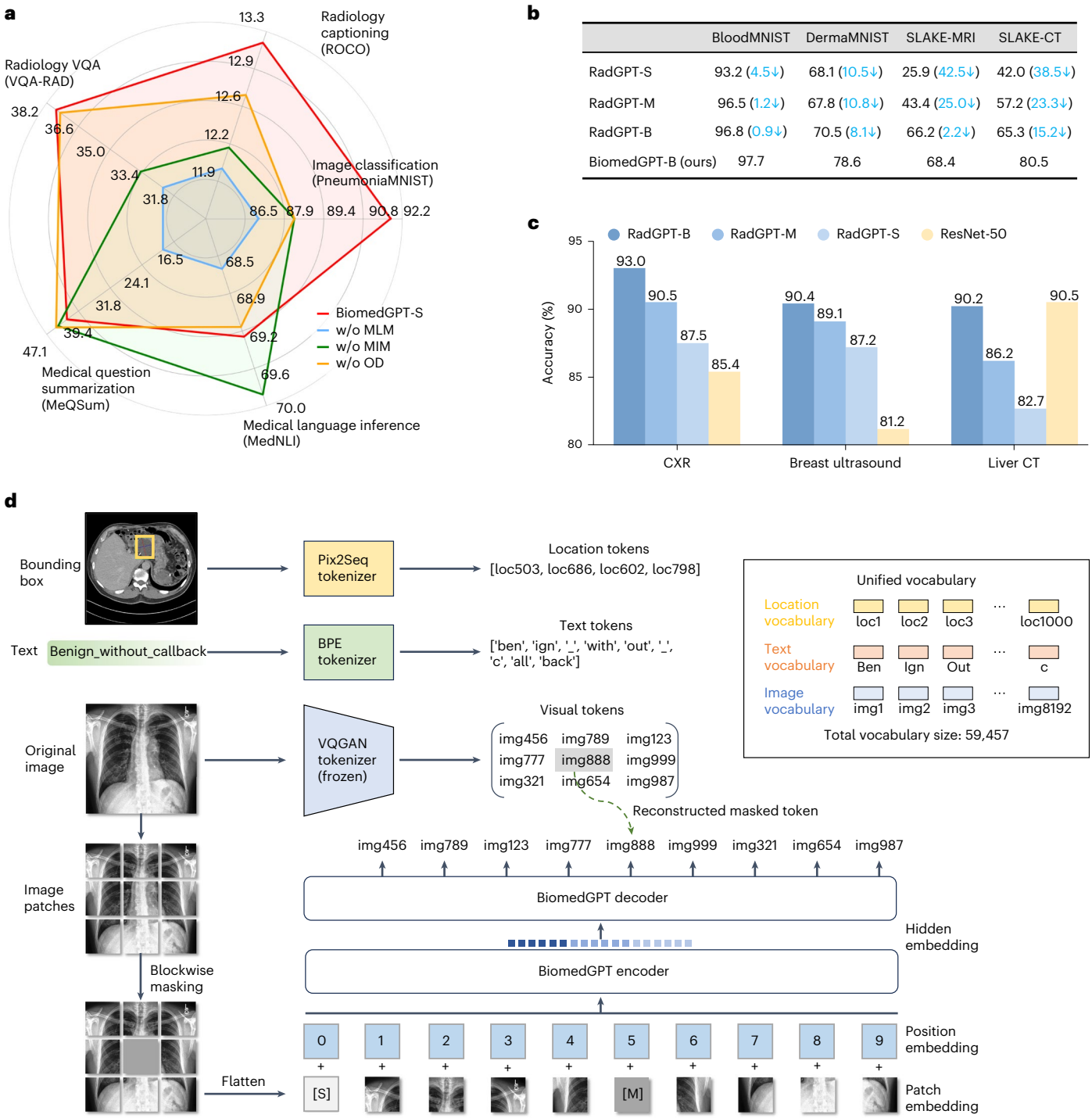


Fig. 6 | Results of the ablation study on the impact of diversity of pretraining datasets and tasks and a graphical demonstration of BiomedGPT's design. a, Performance comparison excluding the specific task. The metrics used are accuracy for radiology VQA, medical language inference and image classification; CIDEr for radiology captioning; and ROUGE-L for medical-question summarization. Pretraining without using masked image modeling, w/o MIM; without using masked language modeling, w/o MLM; without using object detection, w/o OD. **b**, Cross-domain transferability of BiomedGPT across four datasets. RadGPT is a variant of BiomedGPT but was pretrained with radiology-only data. SLAKE-MRI and SLAKE-CT are the modality-specific subsets of the SLAKE data. **c**, In-domain transferability of BiomedGPT across three radiology modalities and datasets. **d**, Description of the unified vocabulary used in BiomedGPT for pretraining and inference. Tokenization of bounding boxes and text was achieved using Pix2Seq and byte-pair encoding (BPE), respectively. There are three types of tokens: location tokens, text tokens and image tokens from frozen pretrained tokenizers, such as VQ-GAN. An illustration of masked image modeling in pretraining, which involves learning representations by reconstructing masked patches, is also shown. [S] and [M] indicate the starting token and masked patch embedding, respectively.

slightly lower at 3.9, versus 4.0 for reference summaries, with no significant difference ($P > 0.05$). BiomedGPT also had a higher correctness rate, with 90.0% of its summaries scoring above 3, compared with 86.0% for the reference impressions. The Wilcoxon rank-sum test showed no significant difference ($P > 0.05$) in average correctness scores between BiomedGPT and the reference summaries, both averaging 4.4 out of 5.

In addition, our analysis found that 6.0% of BiomedGPT-generated summaries contained medically adverse items, categorized as either 'mild' or 'severe', which is identical to the rate observed in the reference impressions. This indicates that BiomedGPT has comparable performance to human experts in summarizing radiology reports, particularly in terms of assessing medical safety. Notably, there was one instance of a 'severe' adverse effect identified in the reference impressions, with no such cases found in the BiomedGPT-generated summaries. The overall score of summaries generated by BiomedGPT closely matches the score of those produced by the reference, with preference scores of 48% for BiomedGPT and 52% for the reference (Fig. 5e). The results of the Sign test ($P > 0.05$) indicate that there is no significant preference for either system, suggesting comparable performance in delivering quality and safety in medical summarization.

Discussion

In this study, we have shown that BiomedGPT can achieve competitive transfer-learning performance across vision, language and multimodal domains by integrating diverse biomedical modalities and tasks within a unified pretraining framework. However, the experimental results also revealed limitations, offering insights for potential improvement.

The development of AI critically depends on the availability of high-quality, annotated data. This requirement poses a unique challenge in the biomedical domain, in which data annotation is expensive, time-consuming and demands extensive domain expertise[54]. Consequently, AI researchers often resort to public datasets, which can compromise data quality. When dealing with multimodal biomedical datasets, particularly image–text pairs, issues become more pronounced: (1) most existing datasets focus primarily on radiology, leading to a substantial modality imbalance; and (2) the scale of images with detailed annotation is still limited in comparison with unlabeled or weakly-labeled biomedical images and accessible biomedical articles from PubMed or PubMed Central. In our study, we considered diverse modalities and ensured that the data scale is sufficient to train high-performance models. As more biomedical data are curated and made open source, we can obtain better visual–semantic mappings (Fig. 6).

Evaluating the quality of generated text presents considerable challenges. Although metrics such as CIDEr and ROUGE-L can measure the agreement between generated content and a gold standard, and are commonly used for model selection to further assess clinical applicability[35], ensuring the factual accuracy of these outputs remains a concern. To address this, recent research has introduced the F_1-RadGraph score[55], which qualitatively assesses the factual correctness and completeness of generated reports. In other domains, such as pathology, similar evaluation metrics are not yet prevalent. We anticipate the emergence of analogous metrics for these domains that draw inspiration from factual-concerned metrics developed in radiology[56]. These would further enhance our ability to measure the factual integrity and overall quality of AI-generated medical content across various biomedical fields.

BiomedGPT is currently adept in processing images and text, and its capabilities could potentially be extended to other types of biomedical data, such as video and time-series or sequential data. For instance, we demonstrated how BiomedGPT can be extended to handle three-dimensional (3D) images by introducing a 3D image encoder into the framework (Extended Data Table 5 and Supplementary Table 4). Nevertheless, these expansions raise concerns about negative transfer, in which learning from additional modalities might inadvertently hamper performance on certain tasks. For instance, our ablation study revealed that excluding image data during pretraining improves performance on language-only downstream tasks (Fig. 6a), highlighting the risk of negative transfer. To mitigate this, we propose exploring controllable learning strategies, such as the mixture of experts[57].

Evidence from our comprehensive analysis (Figs. 3a,b,f and 4a–e,h) indicates a direct correlation between increased model scale and enhanced performance, applicable to both zero-shot predictions and post-fine-tuning. However, scaling brings its own set of challenges, particularly concerning fine-tuning efficiency, training speed and memory requirements. We have tried to address the efficiency challenges of BiomedGPT by exploring prompt tuning, which adds small-scale parameters to condition-frozen models[56]. However, this method incurred large performance degradation (Extended Data Fig. 4b).

Our zero-shot transfer-learning tests (Fig. 4f–h) indicated that BiomedGPT's text-comprehension capabilities, especially in comparison with those of GPT-4V, are not fully established. Two main factors contribute to this limitation: first, the current scale of BiomedGPT, particularly the language backbone, is limited by available resources, although it is expandable. Our preliminary observations indicate that, even if a model has seven billion parameters and effective training, achieving robust zero-shot in-context or text understanding remains challenging in complex medical applications. However, fine-tuning, even with a smaller-scale model such as BiomedGPT, proves to be a promising approach to mitigate risks (Supplementary Fig. 3). Second, the use of a single encoder that handles multiple input types complicates the separation of diverse modality representations, requiring more refined training strategies.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-024-03185-2.

References

1. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
2. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
3. Moody, L. et al. The person-centred care guideline: from principle to practice. *J. Patient Exp.* **5**, 282–288 (2018).
4. Langberg, E. M., Dyhr, L. & Davidsen, A. S. Development of the concept of patient-centredness–a systematic review. *Patient Educ. Couns.* **102**, 1228–1236 (2019).
5. Bates, D. W. et al. Reducing the frequency of errors in medicine using information technology. *J. Am. Med. Inform. Assoc.* **8**, 299–308 (2001).
6. Tu, T. et al. Towards generalist biomedical AI. *NEJM AI* https://doi.org/10.1056/AIoa2300138 (2024).
7. Reed, S. et al. A generalist agent. *Transact. Mach. Learn. Res.* https://openreview.net/pdf?id=1ikK0kHjvj (2022).
8. Driess, D. et al. Palm-e: an embodied multimodal language model. In *Proc. 40th International Conference on Machine Learning* 8469–8488 (JMLR.org, 2023).
9. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems 30* (Neural Information Processing Systems Foundation, 2017).
10. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
11. Touvron, H. et al. Llama: open and efficient foundation language models. Preprint at https://arxiv.org/abs/2302.13971 (2023).
12. Li, C. et al. Llava-med: training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems 36* (Neural Information Processing Systems Foundation, 2024).
13. Wu, C., Zhang, X., Zhang, Y., Wang, Y., & Xie, W. Towards generalist foundation model for radiology. Preprint at https://arxiv.org/abs/2308.02463 (2023).

14. Luo, R. et al. BioGPT: generative pretrained transformer for biomedical text generation and mining. *Brief. Bioinform.* **23**, bbac409 (2022).
15. Zhang, S. et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. Preprint at https://arxiv.org/abs/2303.00915 (2023).
16. Phan, L. N. et al. Scifive: a text-to-text transformer model for biomedical literature. Preprint at https://arxiv.org/abs/2106.03598 (2021).
17. Lau, J. et al. A dataset of clinically generated visual questions and answers about radiology images. *Sci. Data* **5**, 180251 (2018).
18. Liu, B. et al. Slake: a semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *Proc. IEEE International Symposium on Biomedical Imaging (ISBI)* 1650–1654 (Institute of Electrical and Electronics Engineers, 2021).
19. He, X. et al. Towards visual question answering on pathology images. In *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* 708–718 (Association for Computational Linguistics. 2021).
20. Demner-Fushman, D. et al. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Inform. Assoc.* **23**, 304–310 (2016).
21. Johnson, A. E. et al. MIMIC-CXR-JPG — chest radiographs with structured labels. *PhysioNet* **101**, 215–220 (2019).
22. Pavlopoulos, J., Kougia, V., & Androutsopoulos, I. A survey on biomedical image captioning. In *Proc. Second Workshop on Shortcomings in Vision and Language* 26–36 (Association for Computational Linguistics, 2019).
23. Li, P. et al. Self-supervised vision-language pretraining for medial visual question answering. In *Proc. IEEE 20th International Symposium on Biomedical Imaging (ISBI)* 1–5 (Institute of Electrical and Electronics Engineers, 2023).
24. Zhang, X. et al. Pmc-vqa: visual instruction tuning for medical visual question answering. Preprint at https://arxiv.org/abs/2305.10415 (2023).
25. Van Sonsbeek, T. et al. Open-ended medical visual question answering through prefix tuning of language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 726–736 (MICCAI, 2023).
26. Lin, C. Y. Rouge: a package for automatic evaluation of summaries. In *Text Summarization Branches Out* 74–81 (Association for Computational Linguistics, 2004).
27. Banerjee, S. & Lavie, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (eds. Goldstein, J., Lavie, A., Lin, C.-Y. & Voss, C.) 65–72 (Association for Computational Linguistics, 2005).
28. Vedantam, R., Zitnick, C. L. & Parikh, D. Cider: Consensus-based image description evaluation. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)* 4566–4575 (Institute of Electrical and Electronics Engineers, 2015).
29. Jing, B., Xie, P. & Xing, E. On the automatic generation of medical imaging reports. In *Proc. 56th Annual Meeting of the Association for Computational Linguistics* (eds. Gurevych, I. & Miyao, Y.) 2577–2586 (Association for Computational Linguistics, 2017).
30. Chen, Z. et al. Generating radiology reports via memory-driven transformer. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (eds. Webber, B., Cohn, T., He, Y. & Liu, Y.) 1439–1449 (Association for Computational Linguistics, 2020).
31. Liu, F. et al. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 13753–13762 (Institute of Electrical and Electronics Engineers/ Computer Vision Foundation, 2021).
32. Yuan, H. et al. Biobart: pretraining and evaluation of a biomedical generative language model. In *Proc. 21st Workshop on Biomedical Language Processing* (eds. Demner-Fushman, D., Cohen, K. B., Ananiadou, S. & Tsujii, J.) 97–109 (Association for Computational Linguistics, 2022).
33. Van Veen, D. et al. Radadapt: radiology report summarization via lightweight domain adaptation of large language models. In *22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks* (eds. Demner-fushman, D., Ananiadou, S. & Cohen, K.) 449–460 (Association for Computational Linguistics, 2023).
34. Yu, F. et al. Evaluating progress in automatic chest X-ray radiology report generation. *Patterns* **4**, 9 (2023).
35. Van Veen, D. et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.* **30**, 1134–1142 (2024).
36. Jing, B., Xie, P. & Xing, E. On the automatic generation of medical imaging reports. Proc. *56th Annual Meeting of the Association for Computational Linguistics 1* (eds. Gurevych, I. & Miyao, Y.) 2577–2586 (2018).
37. Yang, J. et al. MedMNIST v2 - a large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Sci. Data* **10**, 41 (2023).
38. Jaeger, S. et al. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quant. Imaging Med. Surg.* **4**, 475–477 (2014).
39. Capellán-Martín, D. et al. A lightweight, rapid and efficient deep convolutional network for chest x-ray tuberculosis detection. In *Proc. 2023 IEEE 20th Int. Symp. Biomed. Imaging (ISBI)* 1–5 (IEEE, 2023).
40. Manzari, O. N. et al. Medvit: a robust vision transformer for generalized medical image classification. *Comput. Biol. Med.* **157**, 106791 (2023).
41. Lee, R. S. et al. A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci. Data* **4**, 1–9 (2017).
42. Romanov, A. & Shivade, C. Lessons from natural language inference in the clinical domain. In *Proc. 2018 Conference on Empirical Methods in Natural Language Processing* 1586–1596 (Association for Computational Linguistics, 2018).
43. Gloeckler Ries, L. A. et al. Cancer survival and incidence from the surveillance, epidemiology, and end results (SEER) program. *Oncologist* **8**, 541–552 (2003).
44. Abacha, A. B., & Demner-Fushman, D. On the summarization of consumer health questions. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics* 2228–2234 (2019).
45. Zeng, G. et al. Meddialog: large-scale medical dialogue datasets. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 9241–9250 (Association for Computational Linguistics, 2020).
46. Johnson, A. E. et al. MIMIC-III a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2019).
47. Dubey, S. et al. Using machine learning for healthcare treatment planning. *Front. Artif. Intell.* **6**, 1124182 (2023).
48. Roberts, K. et al. Overview of the TREC 2021 clinical trials track. In *Proc. Thirtieth Text Retrieval Conference* (TREC, 2021).
49. Van Aken, B. et al. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In *Proc. 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* 881–893 (Association for Computational Linguistics, 2021).
50. OpenAI. GPT-4V(ision) system card. *OpenAI* https://openai.com/research/gpt-4v-system-card (2023).

51. Wang, P. et al. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *Proc. Int. Conf. Mach. Learn. PMLR* **162**, 23318–23340 (2022).
52. Hu, X. et al. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In *Proc. 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 4156–4165 (Association for Computing Machinery, 2023).
53. Jeong, J. et al. Multimodal image-text matching improves retrieval-based chest x-ray report generation. In *Proc. Medical Imaging with Deep Learning 227* 978–990 (Proceedings of Machine Learning Research, 2024).
54. Fu, S. et al. Assessment of data quality variability across two EHR systems through a case study of post-surgical complications. In *Proc. AMIA Joint Summits on Translational Science* 196–205 (American Medical Informatics Association, 2022).
55. Delbrouck, J. B. et al. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022* (eds. Goldberg, Y., Kozareva, Z. & Zhang, Y.) 4348–4360 (Association for Computational Linguistics, 2022).
56. Yang, H., Lin, J., Yang, A., Wang, P. & Zhou, C. Prompt tuning for unified multimodal pretrained models. In *Findings of the Association for Computational Linguistics: ACL 2023* (eds. Rogers, A., Boyd-Graber, J. & Okazaki, N.) 402–416 (Association for Computational Linguistics, 2023).
57. Chen, Z. et al. Towards understanding the mixture-of-experts layer in deep learning. *Adv. Neural Inf. Process. Syst.* **35**, 23049–23062 (2022).

[1]Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA, USA. [2]School of Computing, University of Georgia, Athens, GA, USA. [3]Samsung Research America, Mountain View, CA, USA. [4]Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [5]Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, PA, USA. [6]PolicyLab, Children's Hospital of Philadelphia, Philadelphia, PA, USA. [7]Center for Research in Computer Vision, University of Central Florida, Orlando, FL, USA. [8]Department of Computer Science and Engineering, University of California, Santa Cruz, CA, USA. [9]McWilliams School of Biomedical Informatics, UTHealth, Houston, TX, USA. [10]Department of Radiation Oncology, Mayo Clinic, Phoenix, AZ, USA. [11]The Center for Health AI and Synthesis of Evidence (CHASE), University of Pennsylvania, Philadelphia, PA, USA. [12]Penn Institute for Biomedical Informatics (IBI), Philadelphia, PA, USA. [13]Leonard Davis Institute of Health Economics, Philadelphia, PA, USA. [14]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA. [15]Department of Computer Science, Stanford University, Stanford, CA, USA. ✉e-mail: xli60@mgh.harvard.edu; lih319@lehigh.edu; lis221@lehigh.edu

Methods

BiomedGPT is a transformer-based architecture specifically designed for the biomedical field, built on the success of existing unified models for general data. We follow the fundamental principles of a unified model[51]: (1) modality-agnostic, (2) task-agnostic and (3) modality and task comprehensiveness. By discretizing data into patches or tokens, we achieve input–output unification using ideas from ViT[58] and language models[10,11].

BiomedGPT architecture

There are three principal architectures among pretrained foundation models: encoder-only, decoder-only and encoder–decoder. Encoder-only models, such as BERT and its variants[59], primarily use the transformer's encoder to learn representations of input data, and require additional modules, such classification heads or task-specific decoders, during fine-tuning. This architecture may struggle with aligning inputs and outputs across distinctly different modalities, limiting its capability in complex zero-shot prediction or generation tasks. Conversely, decoder-only models, exemplified by GPT[10], rely solely on the transformer's decoder to process raw text inputs. Although proficient in text-based tasks, their architecture is not inherently equipped to handle multiple modalities, often leading to challenges in learning joint representations across diverse data types. This can diminish flexibility and performance in multimodal tasks, particularly in biomedical applications. Therefore, we selected the encoder–decoder architecture to design BiomedGPT, which is more adept at mapping various modalities into a unified semantic representation space, thereby enhancing task handling across a broader spectrum.

BiomedGPT is implemented with a BERT-style encoder[59] over corrupted text and a GPT-style left-to-right autoregressive decoder[10]. All these models rely on the transformer with the popular multi-head attention mechanism (Extended Data Fig. 3a), which allows the model to jointly attend to the information from different representation sub-spaces[60]. To improve the convergence efficiency and stability in the pretraining, we added three normalization operations to each layer: a post-attention Layer Norm (LN)[61], post-first-FFN LN and head-wise scaling within self-attention (Extended Data Fig. 2b), following ref. 62. To encode positional information, we incorporated two sets of absolute position embeddings for both text and images. Rather than merely combining these embeddings with token and patch embeddings, we implemented a decoupling method to separate position correlation (Extended Data Fig. 3b), which could bring unnecessary randomness in the attention and further limit the expressiveness of the model[60]. Furthermore, we also incorporated one-dimensional relative position bias for text and 2D relative position bias for image (Extended Data Fig. 3c), as described in previous works[63,64]. To investigate the performance of BiomedGPT for tasks at different scales, we explicitly designed three scaling models, that is, BiomedGPT-S (33 million parameters), BiomedGPT-M (93 million parameters) and BiomedGPT-B (182 million parameters). The configurations for each model are detailed in Extended Data Figure 2a.

Unifying input–output

To handle diverse modalities without relying on task-specific output structures, we represented them as tokens drawn from a unified and finite vocabulary (Fig. 6d). To achieve this, we used frozen image quantization[65] and object descriptor[66] to discretize images and objects, respectively, on the target side. We encoded text outputs, including object labels and summarizations, using BPE tokens[67]. Specifically, an image with a resolution of 256 × 256 pixels is sparsely encoded into a sequence of 16 × 16 pixels, which correlates strongly with the corresponding patch and can effectively reduce the sequence length of the image representation. The bounding boxes of objects in an image are expressed as sequences of location tokens in the format of integers. We thereby built a unified vocabulary for all tokens of multimodal outputs. The total vocabulary size is 59,457 tokens, including 50,265 language tokens, 1,000 location tokens and 8,192 vision tokens. The number of vision tokens was determined by the variant of the pretrained VQ-GAN models used in BiomedGPT; specifically, we used the variant with a patch size of 8 and vocabulary size of 8,192. During training, we randomly subsampled 196 image patches for pretraining. The maximum model input length is truncated to 512.

Ablation study on modality comprehensiveness. Additional evaluations were conducted to address the query: 'Can the proposed model handle unseen data modalities (for example, images from a new different imaging device like an ultrasound)?' To investigate this, we adjusted our dataset selection for both pretraining and downstream tasks (Supplementary Fig. 2b). Specifically, we used all 3,489 and 6,461 CXR image–text pairs from the SLAKE and IU X-ray datasets, respectively. Additionally, we randomly selected 7,452 images from CheXpert while disabling MLM and OD during pretraining for simplification (Supplementary Fig. 2a). The pretrained BiomedGPT on X-ray modality, denoted as RadGPT-{size}, was then fine-tuned on radiology datasets: CXR, breast ultrasound and liver CT (coronal view). As a comparative baseline, we selected ResNet-50 (ref. 68), which was trained from scratch on these three datasets. We observed impressive in-domain transferability of BiomedGPT from the outcome (Fig. 6c): RadGPT-B outperformed the baseline, achieving 93.0% classification accuracy on the CXR images, a 7.6% improvement. However, for liver CT scans, we had to scale up the model to attain comparable results to the baseline. This highlights the challenges in domain adaptation for medical applications when the pretrained model does not learn diverse medical knowledge.

We further explored the aspect of cross-domain transferability (Fig. 6b). Specifically, we fine-tuned the aforementioned pretrained model, RadGPT, using datasets from other domains, such as blood cell microscopy and dermoscopy, for image classification. Additionally, we selected MRI-only and CT-only image–text pairs from SLAKE and conducted VQA fine-tuning. The results were compared with the benchmark (the original BiomedGPT-B pretrained with all modalities) and were measured in terms of accuracy. We found that cross-modality transfer with our model is feasible, albeit with potentially substantial performance degradation. For example, RadGPT-B exhibited a notable decrease in accuracy compared with the baseline on both the DermaMNIST dataset (dermoscopy), with an 8.1% drop, and the SLAKE-CT VQA dataset, with a more substantial reduction of 15.2%. Notably, we had to double the training epochs as compared with the previous fine-tuning with a pretrained model encompassing all modalities (100 versus 50). Therefore, we conclude that modality comprehensiveness is essential for a generalist biomedical AI model to facilitate efficient knowledge transfer.

Natural language as a task instructor

Multitasking is a key attribute of a unified and generalist model. Following the literature on language models using prompt and instruction learning[10,69,70] and existing unified frameworks to eliminate task-specific modules, we defined each task with a custom instruction, excluding VQA tasks, which are fully specified by their text inputs. BiomedGPT supports abstractions of several tasks, including vision-only, text-only and vision–language, to achieve task comprehensiveness. We provide details of the pretraining tasks and fine-tuning and inference tasks, as well as their corresponding instructions, in the following sections.

Pretraining tasks. We considered two vision-only tasks in the pretraining process: for MIM as well as image infilling, we borrowed the idea of block-wise masking[71] and let the model recover the masked patches in the middle part by generating the corresponding codes (see Fig. 6d).

The corresponding instruction is 'What is the image in the middle part?'. For object detection, the model learns to generate the bounding box of an object with the instruction 'What are the objects in the image?'. For the text-only task, we adopted the commonly used MLM), whose logic is similar to MIM but the instruction is 'What is the complete text of '{Text}'?'. Two types of multimodal tasks were selected, including image captioning with the instruction of 'What does the image describe?' and VQA with the instruction of '{Question}'. The addition of OD for pretraining BiomedGPT serves to enhance visual learning, inspired by ref. 72. The mixture of pretraining tasks is effective, especially for processing multimodal inputs (Fig. 6a).

Fine-tuning and downstream tasks. Besides image captioning and VQA used in pretraining, we covered one more vision-only task and two more text-only tasks. Specifically, we used the instruction 'What does the image describe?' to differentiate image classification. 'What is the summary of text '{Text}'?' and 'Can text1 '{Text1}' imply text2 '{Text2}'?' were exploited for text summarization and natural-language inference, respectively.Notably, BiomedGPT is extendable, allowing for customization of instructions for specific downstream tasks (Fig. 1c and Supplementary Figs. 4–9).

Ablation study on task comprehensiveness. To gain a deeper understanding of the impact of individual pretraining tasks on downstream performance, we implemented an ablation study that excludes either image-only or text-only tasks during pretraining, followed by fine-tuning of the resultant models on five downstream tasks. To ensure a fair comparison, we utilized downstream datasets that were excluded from the pretraining phase: (1) PneumoniaMNIST[36] for image classification; (2) ROCO (https://github.com/razorx89/roco-dataset) for image captioning; (3) VQA-RAD for VQA; (4) MeQSum for text summarization; and (5) MedNLI for text understanding. Moreover, each model was fine-tuned using consistent training receipts across the same datasets.

Owing to the limited computing resources, we performed this study using only BiomedGPT-S. Referring to Supplementary Figure 2c, we used the BiomedGPT-S model, pretrained with all tasks, as the baseline. We observed several empirical phenomena in this ablation study (Fig. 6a): (1) excluding the MIM component resulted in decreased performance in image-centric and multimodal tasks, such as image classification and VQA accuracy. Conversely, text-centric tasks showed improvement. These outcomes indicate that MIM is not crucial for text-only tasks, potentially explaining the enhancements in those areas. (2) When MLM was excluded during pretraining, performance declined across all tasks in downstream evaluation. Text-centric tasks were substantially impacted. These findings underscore the importance of MLM for unified models, even for image-only tasks that require text-token dictionaries for label generation. (3) Excluding object detection during pretraining led to notable performance reductions in tasks such as image classification and radiology captioning. However, changes in performance for other datasets were relatively minor, likely owing to the limited number of object-detection samples and the weak connection to language-only tasks. In summary, our study highlights the importance of task diversity in pretraining for the unified medical AI. Although the exclusion of image-specific tasks might benefit performance on text-only tasks downstream, a varied task regime is essential for maintaining generalization across both unimodal and multimodal applications.

Model pretraining

We adopted sequence-to-sequence (seq2seq) learning[73], which is a commonly used approach for large language models, to train our BiomedGPT. Formally, suppose we are given a sequence of tokens $\mathbf{x}_{i,b}$ as input, where $i = 1, \cdots, I$ indexes the tokens in a data sample and $b = 1, \cdots, B$ indexes a sample in a training batch. Let a model be parametrized by θ. Then we autoregressively train the model by minimizing the loss function L_θ:

$$L_\theta(\mathbf{x}_{1,1}, \cdots, \mathbf{x}_{i,b}) = -\sum_{b=1}^{B} \log \prod_{i=1}^{I} p_\theta(\mathbf{x}_{i,b}|\mathbf{x}_{1,b}, \cdots, \mathbf{x}_{i-1,b}) = -\sum_{b=1}^{B}\sum_{i=1}^{I} \log p_\theta(\mathbf{x}_{i,b}|\mathbf{x}_{<1,b}).$$

In the context of BiomedGPT, $\mathbf{x}$ could refer to both linguistic and visual tokens in the pretraining tasks, including subwords, image codes and location tokens. Specifically, subwords were extracted by a BPE tokenizer, and we masked 15% of the tokens of the subwords in input in the MLM task, because these medical words show relatively high degrees of overlap. For the object-detection task, location tokens are generated following Pix2Seq[66], conditioned on the observed pixel inputs. Data preprocessing was required for quantizing biomedical images using VQ-GAN[67] owing to trivial semantics such as black backgrounds and the need to meet specific input size requirements. Therefore, we first removed the trivial background and cropped the image to the bounding box of the object of interest. We then resized the cropped image to 256 × 256 pixels and fed the center part, with a resolution of 128 × 128 pixels, into the pretrained VQ-GAN to generate the corresponding sparse image codes, which were the target output in masked image modeling task. Vision–language tasks followed the same tokenization flow. For fine-tuning, we also applied seq2seq learning using different datasets and tasks.

To pretrain our BiomedGPT, we used the AdamW[74] optimizer with exponential decay rates for the first and second momentum estimates $\beta_1 = 0.9, \beta_2 = 0.999$, respectively, and a small constant $\varepsilon = 1 \times 10^{-8}$ added to the denominator to improve numerical stability. The peak learning rate is set to 1×10^{-4}, and we applied a linear decay scheduler with a warmup ratio of 0.01 to control the learning rate. For regularization, we set the dropout to 0.1 and used a weight decay of 0.01. To enhance the training process, we used stochastic depth with a rate of 0.1, which was applied to the encoder and decoder, except for convolution blocks. Furthermore, we used a diversified approach in mixing all pretraining data within each batch. This included an assortment of multimodal, text-only, vision-only and object-detection samples. These were used in an 8:2:1:1 ratio to emphasize learning and enhance the interaction between vision and language. In addition, to address the potential feature shift caused by the inherent modality imbalance within the pretraining data, we adopted modality sampling strategies in each pretraining batch to ensure balance. The models were pretrained with 10 NVIDIA A5000 GPUs and mixed precision[75]. Pretraining of the base, medium and small models took approximately 87, 32 and 9 h, respectively. We initialized BiomedGPT with the pretrained OFA model[51] and adapted it to the biomedical domain using our curated multimodal biomedical dataset. Specifically, we continued training from OFA's pretrained checkpoints to align biomedical concepts using diverse modality data through masked modeling, OD and image–text matching (Extended Data Table 3). This approach could reduce computational efficiency as the continued training incorporates general-domain knowledge from OFA, including language-understanding capabilities that are beneficial for question-answering tasks.

Model fine-tuning and inference

Fine-tuning, a form of transfer learning, involves adapting a pretrained model's weights to new data. The practice of fine-tuning pretrained models, a widely acknowledged and highly effective approach in natural-language processing and computer vision, has also found important application in medical AI[76,77]. Unlike most previous biomedical models that necessitate the addition and training of extra components, such as a linear output layer or a decoder, our BiomedGPT model relies solely on fine-tuning the existing structure. The specific instructions used for this fine-tuning procedure mirror those in the pretraining workflow, thereby maintaining consistency and efficiency

in model adaptation. We observed that, in tasks requiring long-context outputs, such as image captioning, the model's performance is influenced by hyperparameters, specifically beam search size and output length constraints (Supplementary Table 6). These findings informed our selection of hyperparameters for fine-tuning, which should be based on data statistics from the training set, such as the maximum length of the target text (Supplementary Table 7). For datasets with an official split, we selected the checkpoint that achieved the highest metric on the validation data for inference during model evaluation (Supplementary Table 7). For datasets lacking an official split, we employed *k*-fold cross-validation, used the checkpoint from the last epoch for inference and reported the mean and s.d.

Similar to existing large language models and multimodal models[28], in inference, we used decoding strategies such as beam search to improve generation quality. However, this approach poses challenges for classification tasks, including unnecessary searching of the entire vocabulary and the possibility of generating invalid labels beyond the closed label set. To tackle these issues, we applied a beam search strategy incorporating a prefix tree (also known as a trie), limiting the number of candidate tokens and resulting in more efficient and accurate decoding. Extended Data Figure 3d demonstrates an example of trie-based beam search; along the path across 'Lipid' and 'breakdown', BiomedGPT sets logits for all invalid tokens ('mechanism' and 'pathway') to $-\infty$ while computing log-probabilities for the target token 'in'. It is worth noting that trie-based search was also applied during the validation phase of the fine-tuning stage for acceleration (approximately 16× increase in speed in our experiments).

Model instruction-tuning and zero-shot prediction

Instruction-tuning was developed to improve the question-understanding capabilities of the pretrained BiomedGPT. Following the data-curation method used for LLaVA-Med[12], we diverged from the traditional VQA approach, in which a pre-built answer set is used during both training and inference. Instead, in our instruction-tuning method, an open-vocabulary setting is used, allowing the model to operate without a predefined set of answers and thereby enabling it to independently determine the most appropriate response during both the training and inference phases.

We summarized experimental settings for each zero-shot trial as follows. In the VQA-RAD zero-shot experiment (Fig. 4), we used the original questions from the dataset as prompts or instructions. For the disease-diagnosis zero-shot experiments (Extended Data Fig. 5b), we used a common prompt template: 'Does the patient have <disease> given the image?'. The evaluation datasets were curated on the basis of the RSNA Pneumonia Detection Challenge (2018) (https://www.rsna.org/rsnai/ai-image-challenge/rsna-pneumonia-detection-challenge-2018) and MedMNIST v2 (images with a resolution of 224 × 224 pixels)[36]. Specific evaluations were conducted across different medical datasets: (1) pneumonia detection involved 1,000 randomly sampled cases from RSNA, including 548 pneumonia and 452 normal cases. (2) Malignant tumor detection used the BreastMNIST dataset, comprising 114 normal or benign cases and 42 malignant cases. (3) Melanoma recognition was based on a subset of DermaMNIST with 223 positive melanoma cases. (4) Drusen recognition utilized a subset of OCTMNIST, featuring 250 positive drusen cases. (5) Cancer tissue identification was assessed on a PathMNIST subset, which included 1,233 colorectal adenocarcinoma epithelium cases, 421 cancer-associated stroma cases, 339 debris cases and 741 normal colon mucosa cases. In TB detection and report generation using two-view CXRs (Extended Data Fig. 5c), we replicated the experimental settings and prompt templates used by Med-PaLM M. Additionally, we incorporated the MIMIC-CXR training set, which includes single-view image–caption pairs, during continual pretraining to ensure a fair comparison with Med-PaLM M. For report generation, we utilized common NLP metrics to align with Med-PaLM M.

Furthermore, we conducted preliminary zero-shot studies on two instruction-tuned large language models, aiming to explore the upper bounds of in-context learning performance using advanced language backbones. We considered the potential integration of these elements into BiomedGPT to enhance reasoning capabilities. However, these models exhibited notable discrepancies when compared with fine-tuned models (Supplementary Fig. 3). These findings suggest that future academic research in medical AI should focus on improving in-context learning abilities and text comprehension, which are crucial for real-world clinical tasks.

Model extension

BiomedGPT was initially developed to process visual (specifically 2D images) and text data. However, the prototype's capabilities could be extended to encompass additional tasks and modalities. For example, we have extended BiomedGPT to include 3D medical imaging classification (Extended Data Table 5 and Supplementary Table 4). This extension involved implementing both pretraining and fine-tuning stages. It requires only integrating a pretrained 3D VQ-GAN for tokenizing 3D images in masked image modeling and adding a learnable 3D visual encoder into the pipeline (Fig. 2a). To further extend the model's capabilities, especially for non-text generation tasks, such as segmentation, introducing additional decoders, such as a mask decoder, is appropriate.

Computing hardware and software

We used Python (version 3.7.4) for all experiments and analyses in the study, which can be replicated using open-source libraries as outlined below. For pretraining, we used ten 24-GB NVIDIA A5000 GPUs configured for multi-GPU training using DistributedDataParallel (DDP) as implemented by the framework PyTorch (version 1.8.1, CUDA 12.2) with the sequence-to-sequence toolkit - fairseq (version 1.0.0). For masked image modeling, we first cropped the middle part of the image and converted it to a sequence of visual tokens based on the pretrained VQ-GAN model (https://heibox.uni-heidelberg.de/d/2e5662443a6b4307b470/). Pillow library (version 9.0.1) was used to read images, which were then converted to the base64 string format using Python. Timm library (version 0.6.12), torchvision (version 0.9.1) and opencv-python (version 4.6.0) were applied for image processing and loading during training. We used the ftfy library (version 6.0.3) to fix potentially broken Unicode for text processing and loading. Einops library (version 0.6.0) was applied for tensor operations in modeling. For model evaluation, we used pycocotools (version 2.0.4) and pycocoevalcap (version 1.2) to calculate the NLP metrics such as ROUGE-L and CIDEr. Other metrics, calculated on the basis of torchmetrics (version 0.11.0). Numpy (version 1.21.5) and Pandas (version 1.3.5), were used in data collection, preprocessing and data analysis.

Evaluation metrics

We used several evaluation metrics to thoroughly assess the capabilities of our BiomedGPT model across different tasks. Accuracy is a primary metric used for evaluating the performance in medical-image classification, VQA and natural-language inference. In addition to accuracy, we also used the F_1 score for the tasks in which class imbalance was considered. The F_1 score is derived as the harmonic mean of precision and recall:

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

For a more convenient comparison with SOTA approaches, we used the weighted F_1 score for VQA. This measure is computed by averaging the F_1 scores across each class, with the individual class scores weighted according to their frequency of occurrence:

$$\text{Weighted } F1 = \sum_{i=1}^{N} \frac{n_i}{N} \times \text{F1}_i,$$

where n_i is the number of instances in class i, N is the total number of instances across all classes and F_{1i} is the F_1 score for class i. Furthermore, we applied the macro-average F_1 score (F_1-macro) in image-classification tasks on the CBIS-DDSM dataset. The F_1-macro score is calculated by determining the F_1 score for each class independently and then averaging these scores across all classes. This approach does not account for class imbalances, treating each class with equal importance:

$$F1-\text{macro}=\frac{1}{N}\times\sum_{i=1}^{N}\text{F1}_i.$$

The higher the accuracy and F_1 score (either weighted- or maro-average), the better performance the model achieves.

ROUGE-L[26] was used to evaluate the quality of the generated text on the image-captioning and text-summarization tasks. Given the candidate C and reference R, let $LCS(C,R)$ be the length of the longest common subsequence, which is determined by using dynamic programming, it can be expressed as:

$$\text{ROUGE}-\text{L}=\frac{(1+\beta^2)R_{LCS}P_{LCS}}{R_{LCS}+\beta^2P_{LCS}},$$

where $R_{LCS}=\frac{LCS(C,R)}{c}$, $R_{LCS}=\frac{LCS(C,R)}{c}$ and $\beta=\frac{P_{LCS}}{R_{LCS}}$. c and r represent the length of the candidate and reference. A higher ROUGE-L score means that the generated text shares more of the same sequences of words as the reference text, which typically indicates better quality in terms of capturing the salient points of the reference. It suggests that the generated text is more similar to the reference summaries that it is being compared with, which is usually desirable in summarization tasks.

In addition to ROUGE-L, we also used METEOR[27] and CIDEr[28] to obtain a more comprehensive evaluation of captioning generation quality. For METEOR, we represented precision and recall as $P=\frac{m}{c}$ and $R=\frac{m}{r}$, where m is the number of common words in the candidate C and the reference R with the number of words of c and r, respectively. METEOR is calculated as follows:

$$\text{METEOR}=(1-p)\frac{PR}{\alpha P+(1-\alpha)R},$$

where p is the penalty factor and is denoted as $p=\gamma(\frac{ch}{m})^{\theta}$, ch is the number of chunks, where a chunk is defined as a set of unigrams that are adjacent in the candidate and reference. α, θ and γ are hyperparameters that are set as 0.1, 3 and 0.5, respectively, in our calculation.

CIDEr is specifically designed to evaluate the quality of image captions. The CIDEr score is calculated using n-gram matching, considering both precision (how many n-grams in the generated caption are also in the reference captions) and recall (how many n-grams in the reference captions are also in the generated caption). It also weighs the n-grams based on their saliency (importance in describing the image) and rarity (uncommonness in the dataset), which helps to emphasize the importance of capturing the most relevant aspects of the image in the caption. CIDEr is obtained by averaging the similarity of different lengths:

$$\text{CIDEr}_n(c,S)=\frac{1}{M}\sum_{i=1}^{M}\frac{\mathbf{g}^n(c)\cdot\mathbf{g}^n(S_i)}{\|\mathbf{g}^n(c)\|\cdot\|\mathbf{g}^n(S_i)\|},$$

where c is a candidate caption, S is set of reference captions, M denotes the number of reference captions and $\mathbf{g}^n(\cdot)$ is an n-gram-based term frequency-inverse document frequency vector. A higher CIDEr score suggests that the generated caption is more accurate and descriptive of the image content, aligning well with human judgments of what the image represents. CIDEr can range from 0 to 100. Typically, human captions tend to score near 90 (ref. 28).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data in this study are publicly available and can be accessed from: IU X-ray and Peir Gross (https://github.com/nlpaueb/bioCaption), MedICat (https://github.com/allenai/medicat), PathVQA (https://huggingface.co/datasets/flaviagiammarino/path-vqa), SLAKE 1.0 (https://www.med-vqa.com/slake/), DeepLesion (https://nihcc.app.box.com/v/DeepLesion), OIA-DDR (https://github.com/nkicsl/OIA), CheXpert- v1.0-small (https://www.kaggle.com/datasets/willarevalo/chexpert-v10-small), CytoImageNet (https://www.kaggle.com/datasets/stanleyhua/cytoimagenet), ISIC 2020 (https://challenge2020.isic-archive.com), Retinal Fundus (https://www.kaggle.com/c/diabetic-retinopathy-detection), MIMIC-III Clinic Notes (https://paperswithcode.com/dataset/hospital-admission-notes-from-mimic-iii), NCBI BioNLP (https://www.ncbi.nlm.nih.gov/research/bionlp/Data/), PubMed abstracts derived from the BLUE benchmark (https://github.com/ncbi-nlp/BLUE_Benchmark), VQA-RAD (https://osf.io/89kps/), CBIS-DDSM (https://www.kaggle.com/datasets/awsaf49/cbis-ddsm-breast-cancer-image-dataset), SZ-CXR and MC-CXR (access can be requested via the contact at http://archive.nlm.nih.gov/repos/chestImages.php), MIMIC-CXR (https://physionet.org/content/mimic-cxr-jpg/2.1.0/), MedNLI (https://physionet.org/content/mednli/1.0.0/), TREC 2022 (https://www.trec-cds.org/2022.html), SEER (https://seer.cancer.gov), MIMIC-III (https://physionet.org/content/mimiciii/1.4/), HealthcareMagic (https://huggingface.co/datasets/UCSD26/medical_dialog), MeQSum (https://huggingface.co/datasets/sumedh/MeQSum), MedMNIST v2 (https://medmnist.com) and ROCO (https://github.com/razorx89/roco-dataset). A randomly sampled subset of RSNA Pneumonia Detection Challenge (2018) was used for zero-shot prediction (https://www.rsna.org/rsnai/ai-image-challenge/rsna-pneumonia-detection-challenge-2018). The MedMNIST-Raw is curated using multiple sources, including NCT-CRC-HE-100K (colon pathology) (https://zenodo.org/records/1214456), HAM10000 (dermoscopy) (https://github.com/ptschandl/HAM10000_dataset), OCT and Chest X-ray (https://data.mendeley.com/datasets/rscbjbr9sj/3), breast ultrasound (https://scholar.cu.edu.eg/Dataset_BUSI.zip), blood cell microscopy (https://data.mendeley.com/datasets/snkd93bnjr/1) and Liver Tumor Segmentation Benchmark (LiTS) (https://competitions.codalab.org/competitions/17094). The VQA data for human evaluation are derived from Medical-Diff-VQA (https://physionet.org/content/medical-diff-vqa/1.0.0/), with the exclusion of questions related to differences, as these require a two-image input. Report generation and summarization samples for human evaluations are extracted from MIMIC-CXR. The instruction-following data used in this article are derived from Pubmed (https://pubmed.ncbi.nlm.nih.gov) following the LLaVA-Med approach (https://github.com/microsoft/LLaVA-Med/blob/main/download_data.sh) and are combined with training sets from PathVQA and SLAKE. We also provided the table with more details of the major datasets in Extended Data Table 2.

Code availability

The pretrained and fine-tuned models, as well as source code for training, inference and data preprocessing, can be accessed at https://github.com/taokz/BiomedGPT.

References

58. Dosovitskiy, A. et al. An image is worth 16×16 words: transformers for image recognition at scale. In *International Conference on Learning Representations*. (2021).

59. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: pretraining of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds. Burstein, J., Doran, C. & Solorio, T.) 4171–4186 (Association for Computational Linguistics, 2019).
60. Ke, G. He, D. & Liu, T. Y. Rethinking positional encoding in language pretraining. In *International Conference on Learning Representations* (ICLR, 2019).
61. Ba, J. L., Kiros, J. R. & Hinton, G.E. Layer normalization. Preprint at https://arxiv.org/abs/1607.06450 (2016)
62. Shleifer, S., Weston, J. & Ott, M., Normformer: Improved transformer pretraining with extra normalization. Preprint at https://arxiv.org/abs/2110.09456 (2021).
63. Dai, Z., Liu, H., Le, Q. V. & Tan, M. Coatnet: marrying convolution and attention for all data sizes. In *Proc. Advances in Neural Information Processing Systems 34 (NeurIPS 2021)* 3965–3977 (Neural Information Processing Systems, 2021).
64. Wang, Z. et al. SimVLM: simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*. (International Conference on Learning Representations, 2022).
65. Esser, P., Rombach, R. & Ommer, B. Taming transformers for high-resolution image synthesis. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 12873–12883 (Institute of Electrical and Electronics Engineers/Computer Vision Foundation, 2021).
66. Chen, T. et al. Pix2seq: a language modeling framework for object detection. In *International Conference on Learning Representations* (International Conference on Learning Representations, 2022).
67. Gage, P. A new algorithm for data compression. *C. Users J.* **12**, 23–38 (1994).
68. He, K. et al. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (Institute of Electrical and Electronics Engineers, 2016).
69. Wei, J. et al. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations* (International Conference on Learning Representations, 2022).
70. Schick, T. & Schütze, H. It's not just size that matters: small language models are also few-shot learners. In *Proc. 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds. Toutanova, K. et al.) 2339-2352 (Association for Computational Linguistics, 2021).
71. Bao, H. et al. BEiT: BERT pretraining of image transformers. In *International Conference on Learning Representations* (International Conference on Learning Representations, 2022).
72. Xu, H. et al. E2E-VLP: end-to-end vision-language pretraining enhanced by visual learning. In *Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (eds. Zong, C. et al.) 503–513 (2021).
73. Sutskever, I., Vinyals, O. & Le, Q.V. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27* (Conference on Neural Information Processing Systems, 2014).
74. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations* (International Conference on Learning Representations, 2019).
75. Micikevicius, P. et al. Mixed precision training. In *International Conference on Learning Representations* (International Conference on Learning Representations, 2018).
76. Raghu, M. et al. Transfusion: understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems 32* (Conference on Neural Information Processing Systems, 2019).
77. Zhou, C. et al. A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT. Preprint at https://arxiv.org/abs/2302.09419 (2023).

Acknowledgements

NSF grant CRII-2246067, NSF POSE: Phase II-2346158 and Lehigh Grant FRGS00011497 supported L.S., K.Z., Z.Y. and Y.L. NIH grant R21EY034179, NSF grants NCS-2319451, MRI-2215789 and IIS-1909879, as well as Lehigh's Accelerator and CORE grants S00010293 and S001250, supported L.H. and R.Z. NIH grants R01HL159183 and RF1AG057892 supported Q.L. NIH grant R03AG078625 supported X.L. NIH grants R01EB19403 and R01LM11934, supported S.F. and H.L. Icons used in Fig. 2 were made by Freepike, surang, Smartline and Blackonion02 at www.flaticon.com.

Author contributions

K.Z. and L.S. designed the study. K.Z., R.Z. and E.A. carried out data collection, data preprocessing, model construction and model validation. J.Y., Z.Y., Y.L. and Z.L. carried out the data analysis benchmarking results. X.C., B.D.D., J.H., C.C., Y.Z., S.F., W.L., T.L., X.L., Y.C., L.H., J.Z., Q.L. and H.L. provided knowledge support and interpreted the findings. H.R. carried out the human evaluation for the generated text from BiomedGPT as well as GPT-4V. L.S. provided knowledge support, interpreted the findings and supervised the study. All authors contributed to manuscript writing and reviewed and approved the final version. L.H., X.L. and L.S. co-supervised the study.

Competing interests

The research was conducted independently of any commercial or financial relationships that could be construed as a potential conflict of interest. Although X.C. is employed by Samsung, the company was not involved in any aspect of this research. The other authors declare no competing interests.

Additional information

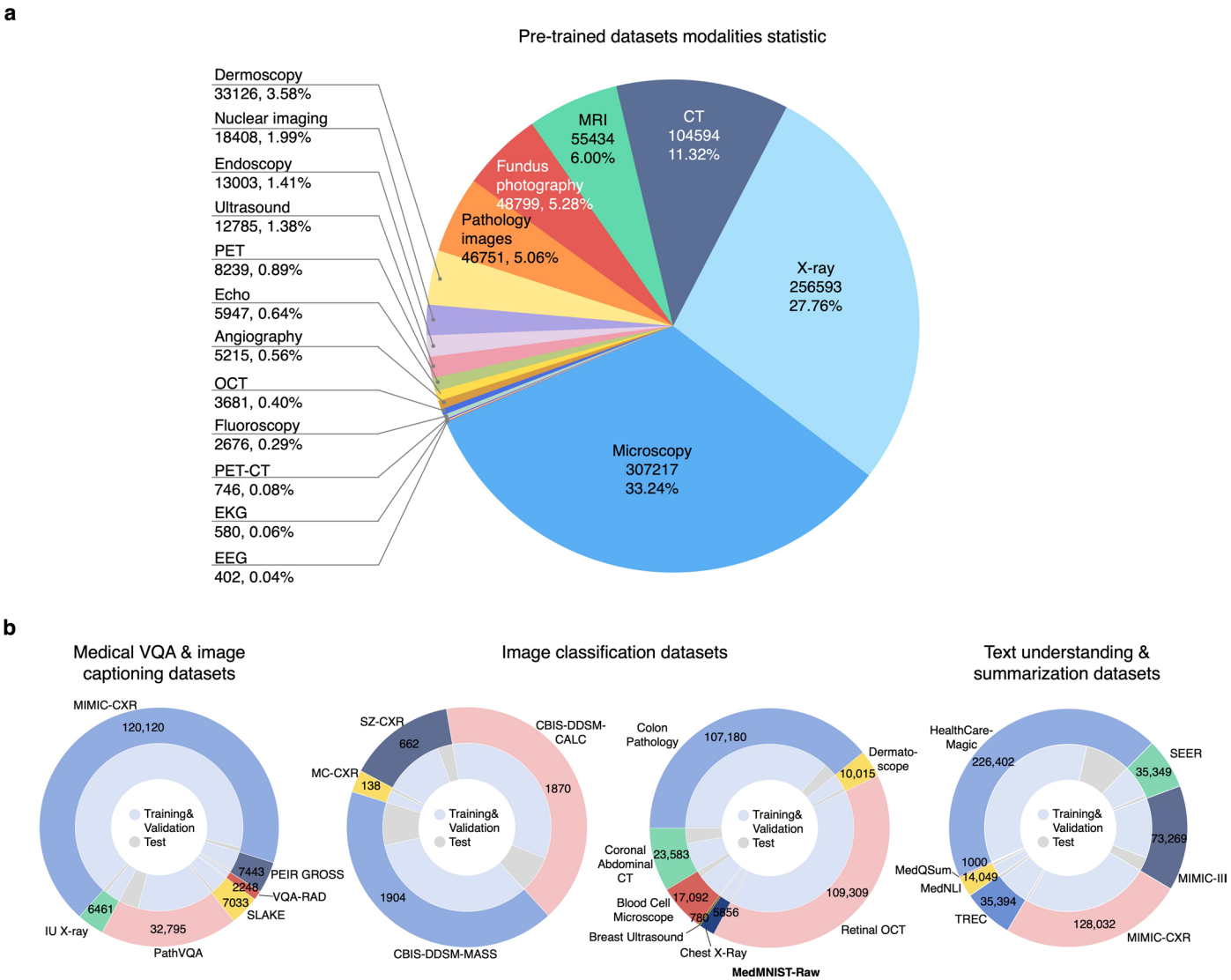
Extended data is available for this paper at https://doi.org/10.1038/s41591-024-03185-2.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-024-03185-2.

Correspondence and requests for materials should be addressed to Xiang Li, Lifang He or Lichao Sun.

Peer review information *Nature Medicine* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Lorenzo Righetto, in collaboration with the *Nature Medicine* team.

https://doi.org/10.1038/s41591-024-03185-2



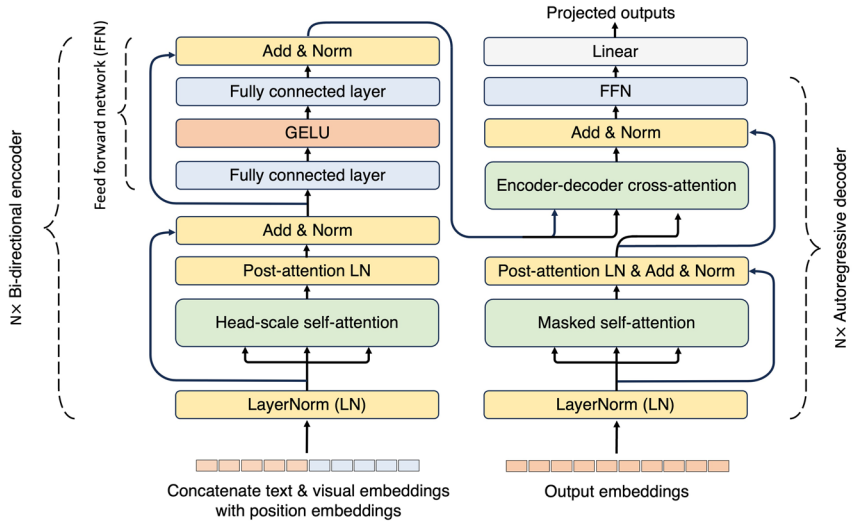
Extended Data Fig. 1 | Statistics of pretraining and fine-tuning datasets. **(a)** Modality distribution of pretraining data used in BiomedGPT. **(b)** For the training and testing splits of datasets used in downstream fine-tuning, we typically follow the format of number of training samples/number of validation samples/number of test samples to detail each dataset. More details of the data split are described in Supplementary Table 7.

https://doi.org/10.1038/s41591-024-03185-2

a

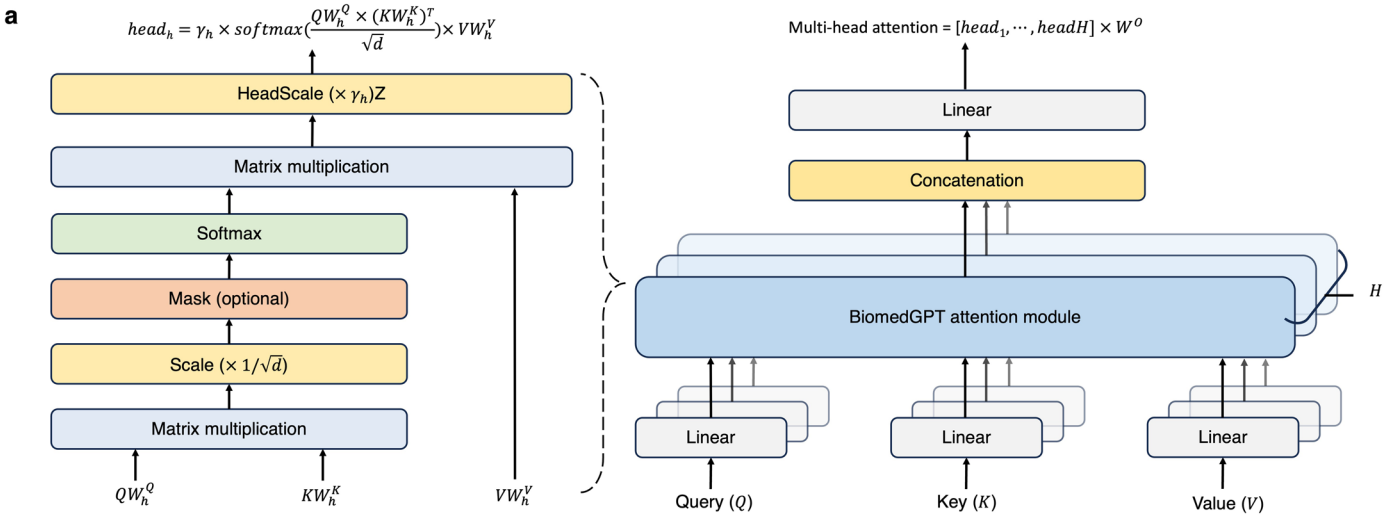
Model scale	#Parameters	Image projection		Representation size		Transformer block		
		Input size	Visual encoder	Hidden	Intermediate	Att. head	#Enc. layer	#Dec. layer
BiomedGPT-S	33 million	256 × 256	ResNet-50	256	1024	4	4	4
BiomedGPT-M	93 million	256 × 256	ResNet-101	512	2048	8	4	4
BiomedGPT-B	182 million	256 × 256	ResNet-101	768	3072	12	6	6

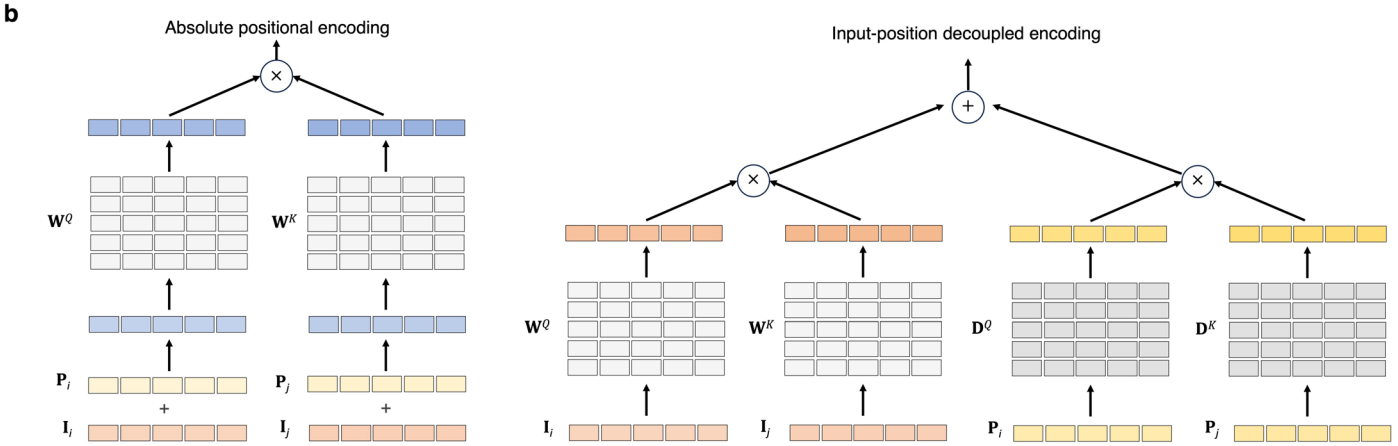
b

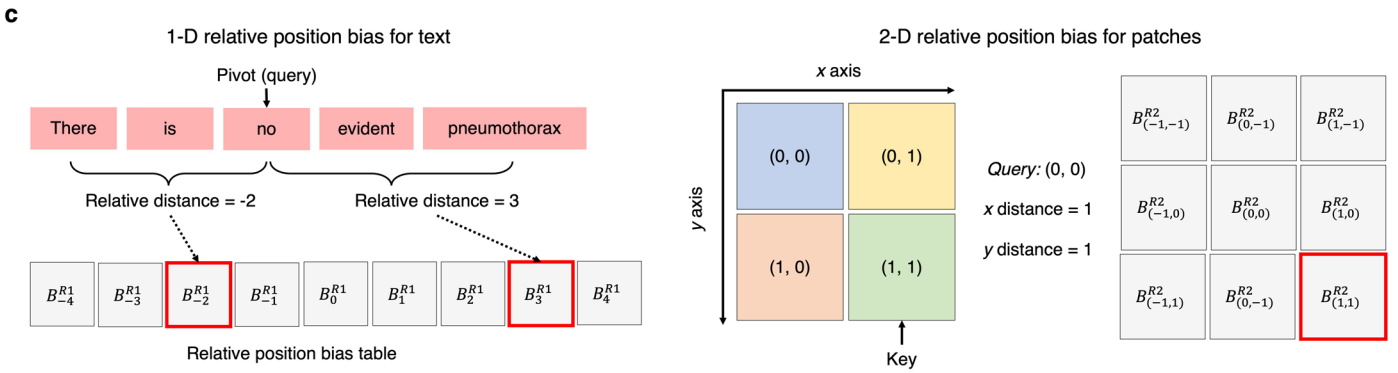


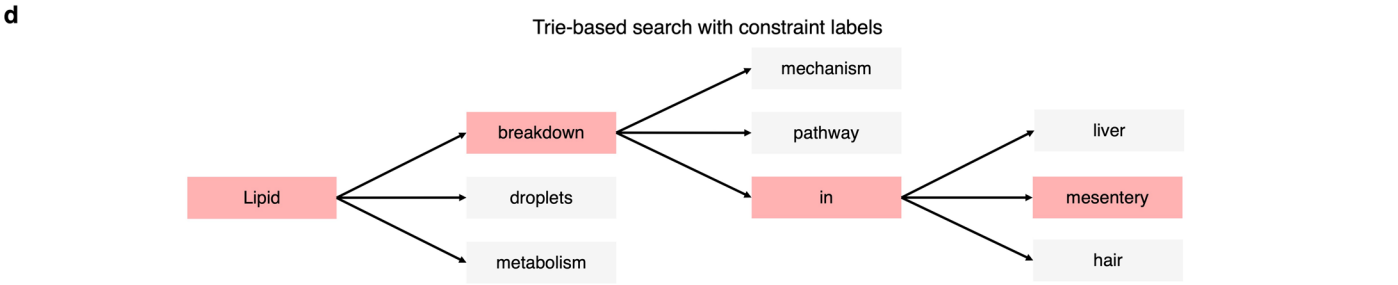
Extended Data Fig. 2 | Overview of BiomedGPT's model configuration and architecture. (a) Detailed model configuration of BiomedGPT. Here, '#' indicates *number of*. **'Att.'**, **'Enc.'** and **'Dec.'** indicate *Attention*, *Encoder* and *Decoder*, respectively. The hidden size is the size of the embeddings and the size of the output of each self-attention and feed-forward layer. The first layer of FFN expands the hidden size to the intermediate size, and the second layer contracts it back to the hidden size. This expansion and contraction allow the network to create more complex representations. During the pretraining phase, image processing involves resizing and cropping the images to varying resolutions, corresponding to the input sizes listed in the table. It should be noted that during fine-tuning and inference stages, the input resolution of BiomedGPT can be flexibly adjusted according to the specific requirements of the task. **(b)** The neural network architecture of BiomedGPT, which includes bidirectional encoder blocks and autoregressive decoder blocks. The number of blocks varies for different model scales.

https://doi.org/10.1038/s41591-024-03185-2







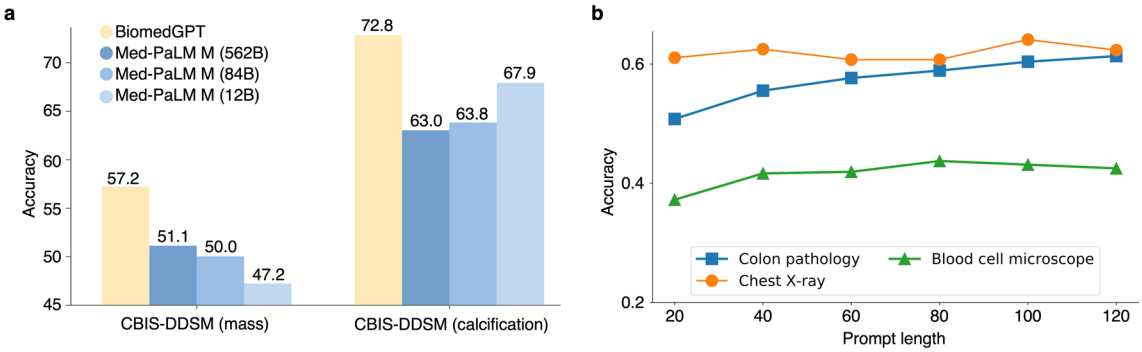


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41591-024-03185-2

Extended Data Fig. 3 | The graphical illustrations of the key components in BiomedGPT. (a) Head-scale multi-head attention module in BiomedGPT. The trainable parameters γ_h is applied prior to the output projection for each head. **(b)** Instead of adding the absolute positional embedding P_i to the input embedding I_i (left), we compute the positional correlation and input correlation separately with different projection matrices and add them together in the self-attention module (right). **(c)** Graphical illustration of relative position bias. Such an inductive bias B_{j-i} is learnable parameter and can be viewed as the embedding of the relative position $j-i$, which is injected into the Query-Key product: $\frac{1}{\sqrt{d}}(I_i W^Q)(P_i W^K) + B_{j-i}$, and shared in all layers. **(d)** An example of trie-based beam search: along the path across 'Lipid' and 'breakdown', BiomedGPT sets logits for all invalid tokens ('mechanism' and 'pathway') to $-\infty$ when computing log-probabilities for the target token 'in'. It is worth noting that trie-based search is also applied during the validation phase of the fine-tuning stage for acceleration (approximately 16× increase in speed in our experiments).

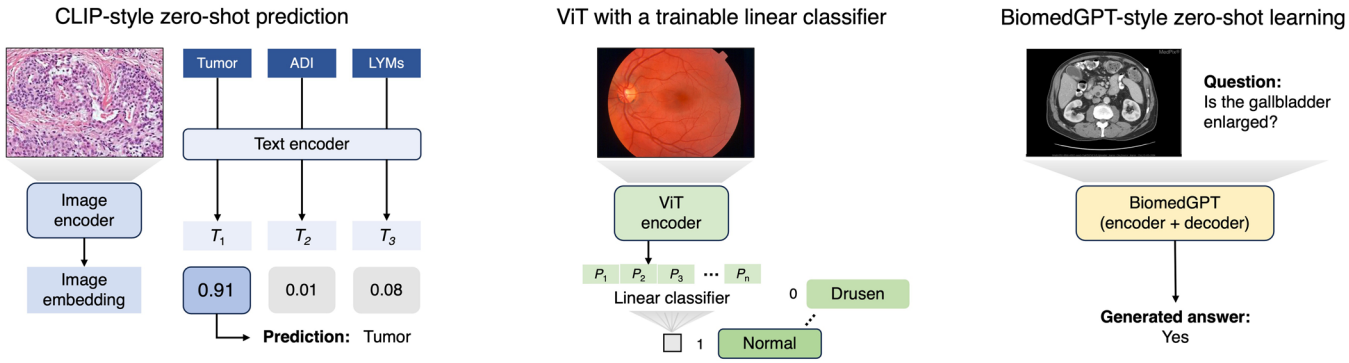
https://doi.org/10.1038/s41591-024-03185-2



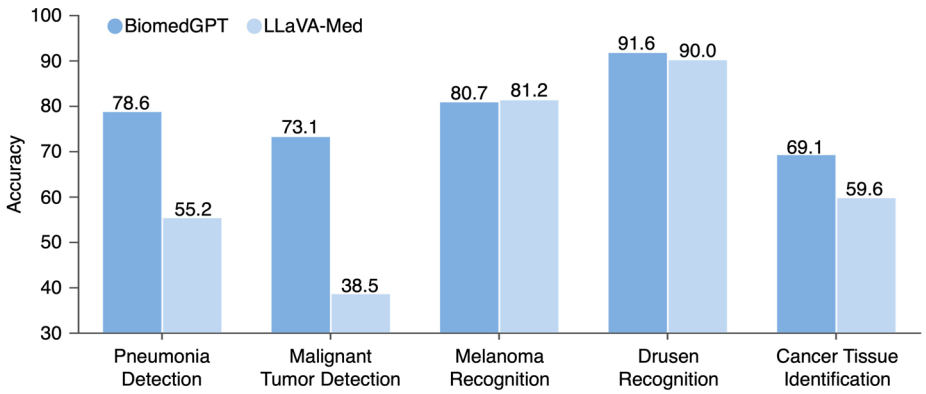
Extended Data Fig. 4 | Comparative Performance of BiomedGPT and Med-PaLM M and the prompt tuning results in Image classification. **(a)** Comparison between BiomedGPT-B and Med-PaLM M on CBIS-DDSM dataset. **(b)** The experimental results of prompt tuning BiomedGPT-B on three image classification datasets. Prompt tuning learns 'soft prompts' or extra model parameters for each task instead of making a task-specific copy of the entire pretrained model for each downstream task and inference must be performed in separate batches. We must mention that the addition of soft prompts is contrary to the design principle of the generalist model. We injected two prompt layers into the encoder and decoder, and varied the prompt length {20, 40, 60, 80, 100, 120} to investigate the performance comparison against full-model fine-tuning. The preliminary results of 'Colon pathology', 'Blood cell microscope', and 'Chest X-ray' were obtained after 100, 512, and 55 training epochs respectively, all with a consistent batch size of 512. We observed that as the prompt length increases, the model performance tends to improve. However, despite an increased number of tuning epochs compared with fine-tuning on the original BiomedGPT (Fig. 3c), the performance after prompt tuning notably lags behind that of model fine-tuning. Specifically, considering only the best results in prompt tuning, there are substantial accuracy reductions of 32.3%, 54.6%, and 32.6% on these three datasets, respectively.

https://doi.org/10.1038/s41591-024-03185-2

a



b



c

Task	Metric	Zero-shot				Fine-tuned	
		BiomedGPT (182 M)	LLaVA-Med (8 B)	Med-PaLM M (12 B)	Med-PaLM M (562 B)	BiomedGPT (182 M)	SOTAs
TB detection	Accuracy	78.3	34.1	87.0	87.7	**89.7**	88.9
Report generation	ROUGE-L	24.4	17.3	27.8	28.5	28.7	**29.6**
	BLEU-4	9.9	4.0	11.5	12.5	**15.4**	13.3
	F1-RadGraph	22.5	9.5	26.7	27.3	**28.0**	24.4
	CIDEr	23.4	0.0	27.6	29.8	**55.2**	49.5

Extended Data Fig. 5 | Additional zero-shot results of BiomedGPT. **(a)** Graphical illustration of zero-shot classification using CLIP-style models, linear probing transfer learning using VIT or BERT-style models, and zero-shot generation of BiomedGPT. Notably, our model can generate the response without providing additional components such as the label candidates for CLIP or linear classifier requiring training for ViT. **(b)** Zero-shot performance on five disease diagnosis tasks. **(c)** BiomedGPT shows competitive zero-shot performance compared with Med-PaLM M with a much smaller model scale. The SOTA fine-tuned model for TB detection is TBLightNet. Note that no single model consistently outperforms the others across all four metrics used in report generation. Here, SOTAs represent the best performance achieved in each specific metric. We fine-tuned our pretrained BiomedGPT-B on MultiMedBench, which Med-PaLM M proposed and used for fine-tuning based on the pretrained PaLM-E. We also attempted to fine-tune LLaVA-Med; however, the time and computational costs were prohibitive due to the large scale of the model and data. Therefore, we reported the results using the pretrained checkpoint of LLaVA-Med.

https://doi.org/10.1038/s41591-024-03185-2

Extended Data Table 1 | Fine-tuned experimental results of BiomedGPT on 25 diverse experiments

Task	Dataset	Domain / Modality	Metric	SOTA		BiomedGPT		
				Model	Result	Small	Medium	Base
Image classification	NCT-CRC-HE-100K	Colon pathology	Accuracy	BiomedCLIP	91.0	94.4	95.8	95.6
	HAM10000	Dermatoscopy		MedViT-L	72.3	66.9	67.6	86.6
	Zhang Lab Data	Retinal OCT		MedViT-L	89.1	84.3	92.9	90.9
		Chest X-Ray		BiomedCLIP	93.0	62.5	94.9	94.9
	Breast Ultrasound	Breast ultrasound		BiomedCLIP	82.2	73.1	73.1	79.5
	Blood Cell Microscope	Blood cell microscope		BiomedCLIP	97.9	82.7	98.5	98.7
	LiTS	Coronal abdominal CT		BiomedCLIP	92.5	56.1	90.6	91.0
	MC-CXR	Chest X-Ray		LightTBNet	88.9	75.9	82.8	89.7
	SZ-CXR	Chest X-Ray			91.0	83.5	97.0	96.2
	CBIS-DDSM	Mass	F1-Macro	Med-PaLM M (562B)	51.1	-	18.7	57.2
		Calcification		Med-PaLM M (12B)	67.9	-	18.9	72.8
Text understanding	MedNLI	Clinic notes	Accuracy	SciFive	85.6	75.8	80.8	83.8
Clinical-Trial Matching	TREC 2022	Clinical trials and patient's medical records	Accuracy	LLaVA-Med	48.7	71.6	75.4	85.2
Treatment suggestions	SEER	Radiation and chemotherapy records	Accuracy	BioGPT	45.9	46.4	49.0	50.0
Mortality prediction	MIMIC-III	ICU admission notes	Accuracy	UMLS-BERT	87.3	77.8	89.2	89.0
Text Summarization	MeQSum	Doctor-patient dialogues	ROUGEL-L	BioBART-L	53.2	42.2	51.3	52.3
	HealthCareMagic	Doctor-patient dialogues	ROUGEL-L	BART-L	44.7	39.8	41.99	42.0
	MIMC-CXR	Radiology report	ROUGEL-L	RadAdapt	44.5	-	-	44.4
			F1-RadGraph	RadAdapt	41.8	-	-	45.1
	MIMIC-III	Radiology report	ROUGEL-L	MedPaLM M (562B)	32.0	-	-	30.7
			F1-RadGraph	MedPaLM M (562B)	34.7	-	-	31.2
Visual question answering	PathVQA	Pathology	Accuracy	CLIP-ViT w/ GPT2	63.6	47.6	49.2	58.1
	VQA-RAD	Radiology		MedVInT-TD	81.6	40.1	69.4	73.2
	SLAKE	Radiology		BiomedCLIP	85.4	69.2	81.6	86.1
Image captioning	IU X-RAY	Chest X-Ray	CIDEr	PPKED	35.1	29.6	31.3	40..1
	PEIR GROSS	Digital camera		CoAttention	32.9	22.0	25.8	122.7
	MIMIC-CXR	Radiology		MedPaLM M (84B)	26.2	-	-	14.7

https://doi.org/10.1038/s41591-024-03185-2

Extended Data Table 2 | Datasets used in BiomedGPT for pretraining, fine-tuning, evaluation with details

Task		Dataset	Availability	Description
Pre-training	Vision & Language	IU X-ray	https://github.com/nlpaueb/bioCaption	A set of chest X-ray images paired with diagnostic reports.
		MediCat	https://github.com/allenai/medicat	A dataset of medical images, captions, and textual references
		PathVQA	https://huggingface.co/datasets/flaviagiammarino/path-vqa	A dataset of question-answer pairs on pathology images.
		PEIR GROSS	https://github.com/nlpaueb/bioCaption	A set of pathology image-caption pairs from PEIR digital library.
		SLAKE	https://www.med-vqa.com/slake/	An English-Chinese bilingual dataset of question-answer pairs.
	Object Detection	DeepLesion	https://nihcc.app.box.com/v/DeepLesion	A dataset with annotated lesions identified on CT images.
		OIA-DDR	https://github.com/nkicsl/OIA	A dataset with annotated fundus images.
	Masked Image Modeling	CheXpert	https://www.kaggle.com/datasets/willarevalo/chexpert-v10-small	A set of chest X-ray images with both frontal and lateral views.
		CytoImageNet	https://www.kaggle.com/datasets/stanleyhua/cytoimagenet	A large-scale dataset of openly-sourced and weakly-labeled microscopy images.
		ISIC	https://challenge2020.isic-archive.com/	Dermoscopic images of unique benign and malignant skin lesions.
		Retinal Fundus	https://www.kaggle.com/c/diabetic-retinopathy-detection	A large set of high-resolution retina images.
	Masked Language Modeling	MIMIC-III Clinic Notes	https://paperswithcode.com/dataset/hospital-admission-notes-from-mimic-iii	A dataset of simulated patient admission notes from MIMIC-III.
		NCBI BioNLP	https://www.ncbi.nlm.nih.gov/research/bionlp/Data/	The corpus contains of annotated PubMed articles.
		PubMed Abstract	https://github.com/ncbi-nlp/BLUE_Benchmark	The corpus consists of annotated PubMed abstracts.
Fine-tuning	Medical VQA	PathVQA	https://huggingface.co/datasets/flaviagiammarino/path-vqa	A dataset of question-answer pairs on pathology images.
		SLAKE	https://www.med-vqa.com/slake/	An English-Chinese bilingual dataset of question-answer pairs.
		VQA-RAD	https://osf.io/89kps/	A dataset of question-answer pairs on radiology images.
	Image Classification	CBIS-DDSM	https://www.kaggle.com/datasets/awsaf49/cbis-ddsm-breast-cancer-image-dataset	A database of scanned film mammography studies.
		MC-CXR	http://archive.nlm.nih.gov/repos/chestImages.php	A dataset of postero-anterior (PA) chest X-rays.
		SZ-CXR	http://archive.nlm.nih.gov/repos/chestImages.php	A dataset of postero-anterior (PA) chest X-rays.
		MedMNIST-Raw	The MedMNIST-Raw is based on multiple datasets: Colon Pathology (NCT-CRC-HE-100K) https://zenodo.org/records/1214456 Dermatoscopy (HAM10000) https://github.com/ptschandl/HAM10000_dataset Retinal OCT & Chest X-ray https://data.mendeley.com/datasets/rscbjbr9sj/3 Breast Ultrasound https://scholar.cu.edu.eg/Dataset_BUSI.zip Blood Cell Microscopy https://data.mendeley.com/datasets/snkd93bnjr/1 Coronal Abdominal CT (LiTS) https://competitions.codalab.org/competitions/17094	Colon Pathology (NCT-CRC-HE-100K): A set of distinct stained histological images patches. HAM10000: A dataset of dermatoscopic images. Dermatoscopy (HAM10000): A large dataset of labeled OCT and Chest X-ray Images Breast Ultrasound: A dataset of breast ultrasound images. Blood Cell Microscopy: A dataset of microscopic peripheral blood cell images. Coronal Abdominal CT (LiTS): Liver Tumor Segmentation Benchmark. A dataset of enhanced abdominal CT scans.
	Image Captioning	IU X-ray	https://github.com/nlpaueb/bioCaption	A set of chest X-ray images paired with diagnostic reports.
		MIMIC-CXR	https://physionet.org/content/mimic-cxr-jpg/2.0.0/	A database of chest X-ray images with free-text reports.
		PEIR GROSS	https://github.com/nlpaueb/bioCaption	A set of pathology image-caption pairs from PEIR digital library.
		ROCO	https://github.com/razorx89/roco-dataset	A large-scale medical and multimodal imaging dataset.
	Text Understanding	MedNLI	https://physionet.org/content/mednli/1.0.0/	A dataset of sentence pairs created by physicians from MIMIC-III clinical notes. For medical language inference.
		TREC2022	https://www.trec-cds.org/2022.html	A dataset of physician-curated sentence pairs from MIMIC-III clinical. For clinical trial matching.
		SEER	https://seer.cancer.gov	A dataset includes cancer information and treatment plans for more than 10,000 patients. For treatment suggestion
		MIMIC-III	https://physionet.org/content/mimiciii/1.4/	A large, de-identified and publicly-available collection of medical records.
	Text Summarization	HealthCareMagic	https://huggingface.co/datasets/UCSD26/medical_dialog	An English-Chinese bilingual dataset of conversations between doctors and patients.
		MedQSum	https://huggingface.co/datasets/sumedh/MeQSum	A dataset of summarized consumer health questions.
		MIMIC-CXR	https://physionet.org/content/mimic-cxr-jpg/2.0.0/	A database of chest X-ray images with free-text reports.
		MIMIC-III	https://physionet.org/content/mimiciii/1.4/	A large, de-identified and publicly-available collection of medical records.
Human Evaluation		Medical-Diff-VQA	https://physionet.org/content/medical-diff-vqa/1.0.0/	A dataset for difference visual question answering on chest X-ray images.
		MIMIC-III	https://physionet.org/content/mimiciii/1.4/	A large, de-identified and publicly-available collection of medical records.
Instruction-following		PubMed articles	https://pubmed.ncbi.nlm.nih.gov	PubMed is a free resource supporting the search and retrieval of biomedical and life sciences literature with the aim of improving health–both globally and personally.
		PathVQA	https://huggingface.co/datasets/flaviagiammarino/path-vqa	A dataset of question-answer pairs on pathology images.
		SLAKE	https://www.med-vqa.com/slake/	An English-Chinese bilingual dataset of question-answer pairs.

https://doi.org/10.1038/s41591-024-03185-2

Extended Data Table 3 | Instructions for pretraining tasks along with the corresponding format of the output

Task	Instructions	The example of output
Masked image modeling	What is the image in the middle part?	<img111> <img222> <img333> ... <img999>
Masked language modeling	What is the complete text of "Effect of <mask> on cultured fibroblasts" ?	Effect of **chloroquine** on cultured fibroblasts
Object detection	What are the objects in the image?	<loc111> <123> <loc789> <loc567> **chest** <loc222> <333> <loc666> <loc999> **kidney**
Image captioning	What does the image describe?	Interval placement of endotracheal tube and nasogastric tube in standard position.
Visual question answering	{**Question**}	{**Answer**}

Here, *<img>* represents the image token derived from VQ-GAN's vocabulary. *<loc>* represents the location token. The instruction for the VQA task is the question itself from the dataset.

https://doi.org/10.1038/s41591-024-03185-2

Extended Data Table 4 | Description of the question types for human evaluation

Type	Explanation
Modality recognition	The specific imaging modality, such as CT, MRI, or others.
Structural identification	The specific anatomical landmarks or structures within the captured images.
Lesion & abnormality detection	The identification of anomalous patterns or aberrations
Disease diagnosis	Specific disease or medical conditions based on imaging manifestations
Size & extent assessment	The dimensions and spread of a lesion or abnormality.
Spatial relationships	The relative positioning or orientation of imaged structures.
Image technical details	The nuances of the imaging process itself, such as contrast utilization or image orientation

Description of the question types in the selected VQA-RAD data samples, which are used for the evaluation of zero-shot learning performance.

https://doi.org/10.1038/s41591-024-03185-2

Extended Data Table 5 | 3D medical image classification performance

Model	Parameters	AIBL		MIRIAD		LIDC	
		Accuracy	F1-macro	Accuracy	F1-macro	Accuracy	F1-macro
BiomedGPT-B-3D	182 M	88.6	77.8	84.7	83.0	92.9	92.1
BiomedGPT-M-3D	93 M	84.7	72.1	80.0	77.5	89.9	88.9
MedicalNet-101	99 M	81.8	66.8	70.6	65.5	89.9	88.5
MedicalNet-152	152 M	85.7	72.6	78.2	75.9	90.9	89.5
COVID-ViT	78 M	64.4	51.0	33.8	33.3	91.9	90.8
Uni4Eye	340 M	69.7	55.8	64.7	59.5	84.9	82.8

3D medical image classification performance in terms of accuracy and F1-Macro. (Details of data and training are described in Supplementary Table 4).

nature portfolio

Corresponding author(s): Xiang Li, Lifang He, Lichao Sun

Last updated by author(s): Jul 1, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Scripts for data collection and preparation was written in Python (version 3.7.4) using the libraries Numpy (version 1.21.5), Pandas (version 1.3.5), and Pillow (version 9.0.1)
Data analysis	Codes for data analysis was written in Python (version 3.7.4). The following Python libraries were used for analysis: PyTorch (version 1.8.1, CUDA 12.2), fairseq (version 1.0.0), Timm (version 0.6.12), torchvision (version 0.9.1), opencv-python (version 4.6.0), ftfy (version 6.0.3), einops (version 0.6.0). pycocotools (version 2.0.4) and pycocoevalcap (version 1.2) to calculate the natural language processing (NLP) metrics such as ROUGE-L and CIDEr. Other metrics and analysis are performed based on torchmetrics (version 0.11.0), Numpy (version 1.21.5) and Pandas (version 1.3.5).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data in this study are publicly available and can be accessed from: IU X-ray and PEIR GROSS (https://github.com/nlpaueb/bioCaption), MedICat (https://github.com/allenai/medicat), PathVQA (https://github.com/UCSD-AI4H/ PathVQA), SLAKE 1.0 (https://www.med-vqa.com/slake/), DeepLesion (https://nihcc.app.box.com/v/DeepLesion), OIA-DDR (https://github.com/nkicsl/OIA), CheXpert- v1.0-small (https://www.kaggle.com/datasets/willarevalo/chexpert-v10-small), CytoImageNet (https://www.kaggle.com/datasets/stanleyhua/cytoimagenet), ISIC 2020 (https://challenge2020.isic-archive.com), Retinal Fundus (https://www.kaggle.com/ c/diabetic-retinopathy-detection), MIMIC-III Clinic Notes (https://paperswithcode.com/dataset/hospital-admission-notes-from-mimic-iii), NCBI BioNLP (https://www.ncbi.nlm.nih.gov/research/bionlp/Data/), PubMed Abstracts are derived from BLUE benchmark (https://github.com/ncbi-nlp/BLUE_Benchmark), VQA-RAD (https://osf.io/89kps/), CBIS-DDSM (https://www.kaggle.com/datasets/awsaf49/cbis-ddsm-breast-cancer-image-dataset), SZ-CXR and MC-CXR can be requested via the contact on (http://archive.nlm.nih.gov/repos/chestImages.php), MIMIC-CXR (https://physionet.org/content/mimic-cxr-jpg/2.1.0/), MedNLI (https://physionet.org/content/mednli/1.0.0/), TREC2022 (https://www.trec-cds.org/2022.html), SEER (https://seer.cancer.gov), MIMIC-III (https://physionet.org/content/mimiciii/1.4/), HealthcareMagic (https://github.com/UCSD-AI4H/Medical-Dialogue-System), MeQSum (https://huggingface.co/datasets/sumedh/MeQSum), MedMNIST v2 (https://medmnist.com) , ROCO (https://github.com/razorx89/roco-dataset), a randomly sampled subset of RSNA Pneumonia Detection Challenge (2018) used for zero-shot prediction (https://www.rsna.org/rsnai/ai-image-challenge/rsna-pneumonia-detection-challenge-2018). The MedMNIST-Raw is curated based on multiple sources including NCT-CRC-HE-100K (colon pathology) (https://zenodo.org/records/1214456), HAM10000 (dermatoscope) (https://github.com/ptschandl/HAM10000_dataset), OCT & Chest X-ray (https://data.mendeley.com/datasets/rscbjbr9sj/3), breast ultrasound (https://scholar.cu.edu.eg/Dataset_BUSI.zip), blood cell microscopy (https://data.mendeley.com/datasets/snkd93bnjr/1), Liver Tumor Segmentation Benchmark (LiTS) (https://competitions.codalab.org/competitions/17094). The VQA data for human evaluation are derived from Medical-Diff-VQA (https://physionet.org/content/medical-diff-vqa/1.0.0/), with the exclusion of questions related to differences, as these require a two-image input. Report generation and summarization samples for human evaluations are extracted from MIMIC-CXR. The instruction-following data used in this article is derived from Pubmed (https://pubmed.ncbi.nlm.nih.gov) following LLaVA-Med (https://github.com/microsoft/LLaVA-Med/blob/main/download_data.sh) and is combined with training sets from PathVQA and SLAKE.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For pre-training, 237,621 medical images paired with natural language, 13,673 diabetic retinopathy images with object bounding boxes, 32,735 CT images with object bounding boxes, 5,126 retinal fundus images, 33,126 skin lesion images, 224,315 chest radiology images, and 300,000 microscopy images, and about 182 million sentences from medical articles and clinical notes were used in this study to ensure an adequate representation of medical data under investigation. We did not strategically select specific numbers of samples for each modality. Initially, we sought to gather as much data as possible and subsequently aimed to cover a wide range of modalities and tasks (especially for captioning, object detection and VQA, which require limited labeled data). Our approach was primarily driven by the availability of data, focusing on achieving extensive coverage to enhance the model's versatility within our resource constraints.

Data exclusions	For pre-training data curation, we excluded approximately 590K cases from CytoImageNet and retained 300K cases. This exclusion was performed to prevent the dominance of microscopy images in the pre-training data, which could hinder the model's ability to learn representations of other modalities.
Replication	We confirm that all experimental findings can be reproduced with our source code provided.
Randomization	For treatment suggestion and clinical trial matching tasks, we employed 10-fold cross validation method, while for each fold 80% cases were randomly sampled as the training data, the remaining 20% as the test data. For other datasets, we adhered to the official training/validation/test splits provided by the data owners. These official splits are widely adopted by the research community, allowing us to easily compare our model's performance with others. This approach avoids the need to rerun all experiments of baselines, particularly when some models are not accessible for producing results with covariance.
Blinding	In our human evaluations (VQA and report summarization), the rater were blind to the source of the response (model or gold standard from phycisian).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A