Assessing large multimodal models for one-shot learning and interpretability in biomedical image classification

Wenpin Hou[1,†], **Yilong Qu**[2], **and Zhicheng Ji**[2,†]

[1]Department of Biostatistics, The Mailman School of Public Health, Columbia University, New York City, NY, USA
[2]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA.
[†]Corresponding author. E-mail: wh2526@cumc.columbia.edu; zhicheng.ji@duke.edu

ABSTRACT

Image classification plays a pivotal role in analyzing biomedical images, serving as a cornerstone for both biological research and clinical diagnostics. We demonstrate that large multimodal models (LMMs), like GPT-4, excel in one-shot learning, generalization, interpretability, and text-driven image classification across diverse biomedical tasks. These tasks include the classification of tissues, cell types, cellular states, and disease status. LMMs stand out from traditional single-modal classification approaches, which often require large training datasets and offer limited interpretability.

Introductions

Large Multimodal Models (LMMs)[1–5] are an advanced extension of Large Language Models (LLMs), incorporating multi-sensory capabilities to process and integrate diverse data modalities such as text, images, and audio. The ability to incorporate natural language makes LMMs distinct from single-modal methods that are built and trained solely on images. Although LMMs may not fully understand concepts in natural language, they can link these concepts with features in images and use them as tools to efficiently capture and convey information. In contrast, single-modal methods require representing features in images with a large number of parameters, and the training process often involves updating millions of parameters in a complex neural network.

We hypothesize that LMMs outperform conventional single-modal deep-learning methods in biomedical imaging classification, offering advances in *post hoc* interpretability and higher accuracy. Image classification is a classic application of deep learning methods, but the lack of interpretability has hindered the widespread application of these "black boxes" in the biomedical field. In contrast, LMMs, inheriting language capabilities from LLMs, can provide reasoning steps for image classification in plain language. This *post hoc* interpretability is crucial for building trust, ensuring transparency, and identifying errors in clinical decision-making. To test this *post hoc* interpretability, we design one-shot learning tasks in this study to evaluate and compare the performance of LMMs and single-modal models. "One-shot," rather than "few-shot," was chosen to highlight the application for rare diseases, where collecting, annotating, and curating training samples can be time-consuming, labor-intensive, or even inaccessible. Compared to single-modal methods, which often require vast amounts of training images, pretrained LMMs have demonstrated superior performance in one-shot learning[3,6], where only one example is used in the training process. This approach eliminates the need for large training datasets, significantly reducing the complexity of applying these methods. Additionally, biomedical images generated by different labs, under varying conditions, or with differing experimental procedures, are often not directly comparable, leading to discrepancies between training and testing datasets. LMMs could be better suited to handle such discrepancies, as concepts expressed in natural language are more abstract and can generalize across images to capture common features of interest. In contrast, single-modal methods are more vulnerable to the challenges posed by dataset shifts. Finally, LMMs support in-context learning, where users can directly provide training examples via prompt messages without needing to retrain or fine-tune the model. Since several commercial LMMs offer user-friendly interfaces that require no programming skills, they have the potential to significantly lower the barrier to performing biomedical image classification tasks, improving health equity and benefiting hard-to-reach populations and resource-limited areas. By contrast, implementing single-modal methods often requires substantial programming skills and deep learning expertise.

In this study, we systematically benchmarked the performance of several leading commercial LMMs, including GPT-4o, Claude 3, and Gemini 3.5, across six types of biomedical image classification tasks. We included only commercial LMMs with

user-friendly interfaces to emphasize their accessibility to healthcare providers without programming expertise. Our evaluation focused on the one-shot learning ability and interpretability of LMMs, which we believe are the aspects that most clearly distinguish LMMs from single-modal methods. To test the generalizability of LMMs, we introduced discrepancies between the training and testing data by applying various transformations to the training images. For comparison, we repeated all analyses using ten single-modal image classification methods. Our findings show that LMMs substantially outperform single-modal methods in one-shot learning, interpretability, and generalizability, highlighting the unique advantages of LMMs in biomedical image classification and offering new insights not reported in previous benchmark studies[6–8].

Results

A prompt strategy for image classification

We developed a prompt strategy for an LMM to perform image classification (Figure 1a). For a binary classification task, a training image array with two rows and two columns is created. The two rows represent two image categories, and the first column labels these as "row 1" and "row 2." The second column contains two training images, one for each category. The training image array can easily be created using PowerPoint, Keynote, or similar software. A dictionary of tissues, cell types, or disease conditions corresponding to the two rows is stored on the user's local system and is hidden from the LMM. When presented with a test image, the LMM is asked to identify which row in the training image array most closely resembles the test image. The user can then refer to the locally stored dictionary to determine the actual category.

This strategy offers several advantages. First, additional rows and columns can be added to represent more image categories or include more training images per category, thus bypassing the maximum input file limits imposed by many LMMs. Second, compared to providing separate image files for each category, the reference image array simplifies communication for both the user and the LMM. Third, sensitive information, such as patient data, can be stored in a separate dictionary for enhanced confidentiality, which is especially important in clinical settings. Finally, this design minimizes the chance that the LMM will rely on prior knowledge of image category names (e.g., cell type names), ensuring a fair comparison with single-modal methods. Figure 1b illustrates a real example where GPT-4 was queried to classify images of normal and senescent cells using this prompt strategy.

To evaluate the interpretability of LMMs, we designed a text-based image classification prompt that uses the LMMs' natural language output from a previous conversation as training information in a completely new conversation (Figure 1c-d). First, the LMM was asked to summarize the main differences between the two image categories in the training image array (Figure 1c). Then, in a new conversation, the LMM was queried to identify the category of a test image using only the summarized text, without access to the original training images (Figure 1d).

LMMs outperform single-modal methods in one-shot learning

We systematically evaluated the performance of three commercial LMMs (GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro) for in-context one-shot learning, with one training image provided for each image category. Additionally, we included ten single-modal methods, AlexNet[9], DenseNet[10], EfficientNetV2[11], GoogLeNet[12], Inception V3[13], ResNet[14], ResNeXt[15], SwinTransformer[16], VGG[17], and VisionTransformer[18], which were benchmarked on the same sets of training and testing images.

We designed three biology image classification tasks and three medical image classification tasks (Figure 2). The biology image tasks include classifying artery and tibial nerve tissues using hematoxylin and eosin stain (HE) images, classifying microglial (BV-2 cell line) and neuroblastoma (SH-SY5Y cell line) cells using light microscopy images, and classifying normal and senescent cells using DAPI fluorescent stain images. The medical image tasks include classifying normal retinas and retinas with diabetic macular edema (DME) using optical coherence tomography (OCT) images, classifying normal brains and brains with tumors using magnetic resonance imaging (MRI), and classifying normal lungs and lungs with tumors using chest computed tomography (CT) scans.

Figure 3a shows the accuracy, defined as the percentage of test images correctly categorized (Methods), of each method across the classification tasks. The accuracy and responses of the three LMMs are detailed in Supplementary Table 1. GPT-4o demonstrates the best performance, achieving nearly 100% accuracy in all scenarios. The other two LMMs, Claude 3.5 and Gemini 1.5, also show strong performance in most scenarios but do not match the performance of GPT-4o. The ten single-modal methods perform worse than the three LMMs. For instance, SwinTransformer, the best-performing single-modal method, performs significantly worse than GPT-4o in the three medical image tasks, despite achieving near-perfect performance in the biology image tasks. A potential explanation is that the differences between the two image categories in the medical tasks are smaller and more subtle compared to those in the biology tasks. As a result, single-modal methods may require more training data to effectively capture such differences. In contrast, GPT-4o is able to capture these differences using only one training image per category.

LMMs outperform single-modal methods in generalizability

We further tested the models' generalizability, which refers to their capacity to apply what they have learned from training images to systematically different testing images. While keeping the testing images unchanged, we introduced heterogeneity between the training and testing images by applying various transformations to the training images or by collecting training images from different studies (Methods, Figure 2).

Figure 3b shows the performance of all models when the training images and testing images are heterogeneous. Figure 3c further compares the performance of each model when the training images are either homogeneous or heterogeneous relative to the testing images. While the performance of all models decreases when the training and testing images are heterogeneous, the performance of LMMs declines much less than that of most single-modal models. Specifically, GPT-4o experiences only a slight performance drop, significantly outperforming all single-modal methods in nearly all scenarios. The performance of SwinTransformer drops considerably in biology image tasks compared to its near-perfect performance when the training and testing images are homogeneous. These results suggest that LMMs are more generalizable than single-modal methods and that LMMs can still reliably perform image classification tasks even when the training and testing images are heterogeneous.

GPT-4o demonstrates interpretability

Beyond accuracy, we also evaluated the interpretability of GPT-4o, the best-performing LMM. We summarized the keywords from GPT-4o's responses describing differences between the two image categories, and their enrichment is visualized as word clouds (Figure 4, Methods). The keywords align well with the features distinguishing the two image categories. For example, "dense" and "homogeneous" describe the image features of an artery, while "round" and "circular" describe those of a tibial nerve in tissue classification. In disease diagnostics, "uniform" and "smooth" describe the features of a normal retina, while "peak" and "irregular" describe those of DME. These results highlight GPT-4o's interpretability in explaining the reasoning behind image classification, an ability not seen in single-modal methods. Additionally, the keywords provided by GPT-4o remain highly consistent regardless of the heterogeneity between the training and testing data, suggesting that GPT-4o can capture the most relevant information in image classification tasks, which may explain its superior generalizability. In contrast, single-modal methods can be misled by irrelevant information in the training data that does not transfer effectively to the testing data.

Finally, we performed the text-based image classification method described in the previous section using GPT-4o. The natural language that GPT-4o itself used to summarize the differences between image categories served as the sole information carrier for the training images. The performance, labeled as "GPT-4o (text)" in Figure 3, is only slightly worse than GPT-4o's image-based classification ("GPT-4o (image)") and outperforms all single-modal methods in almost all scenarios, achieving a reliable accuracy of at least 89% in all tasks. GPT-4o's text-based classification also demonstrates superior generalizability, similar to its image-based classification (Figure 3c). These results highlight GPT-4o's ability to interpret imaging features as natural language concepts, which can then be used to classify test images, even in a one-shot learning setting. Additionally, text-based classification offers an alternative approach to image-based classification, particularly when sharing training images is difficult due to technical or privacy concerns.

Conclusions

Our study illustrates that LMMs, such as GPT-4o, are able to leverage natural language as a foundational component for reasoning in image classification. This feature substantially enhances LMMs' capabilities in one-shot learning, generalizability, interpretability, and text-based image classification, surpassing single-modal methods. These findings indicate that LMMs, which integrate natural language with visual data, have potential advantages over single-modal methods that rely solely on images. This insight could inform future research directions in computer vision, highlighting the benefits of combining linguistic and visual data. Moreover, the text-based classification approach suggests the possibility of developing new image analysis methods that utilize language-based communication between artificial intelligence systems. This approach could offer a more comprehensive way of interpreting and interacting with complex visual data, potentially enhancing the utility and applicability of AI in various domains by bridging the gap between human interpretative skills and machine efficiency.

Methods

Imaging data

Biology image classification tasks

For tissue classification, H&E images of artery and tibial nerve tissues were downloaded from the GTEx portal[19]. To create a scenario where training and testing images are homogeneous, we collected fifteen images of artery from different randomly chosen samples and fifteen images of tibial nerve from different randomly chosen samples. We randomly zoomed in on each of the original image provided by GTEx to display a part of the tissue, and all images have the same zoom level. We created five

sets of training images, each set containing one artery image and one tibial nerve image. All the remaining images were used for testing. To create a scenario where the training and testing images are heterogeneous, we further zoomed in on each image of the same training set to display a smaller region of the tissue. The testing set remained the same.

For cell type classification, light microscopy images of BV-2 and SH-SY5Y cell lines were downloaded from the LIVECell dataset[20]. To create a scenario where training and testing images are homogeneous, fifteen images from each of the two cell lines were randomly chosen. Five images from each cell line were used to create five sets of training images, and the remaining images were used for testing. To create a scenario where the training and testing images are heterogeneous, we downloaded light microscopy images of BV-2 and SH-SY5Y from two additional studies[21,22]. For each image, we randomly cropped five different regions of the original image to obtain five new images for training. The testing set remained the same.

For cell state classification, DAPI fluorescence images were downloaded from a previous study[23]. To create a scenario where training and testing images are homogeneous, for each cell state of normal or senescence, fifteen images of single cells were randomly cropped from the original images in the study. Five images from each cell state were used to create five sets of training images, and the remaining images were used for testing. To create a scenario where the training and testing images are heterogeneous, we downloaded DAPI fluorescence images from another study[24]. For each cell state, five images of single cells were randomly cropped from the original images and five new sets of training images were created. The testing set remained the same.

Medical image classification tasks

For retina OCT images, OCT images were downloaded from a previous study[25]. To create a scenario where training and testing images are homogeneous, fifteen images were randomly selected for each disease status. Five images for each disease status were used for training and the remaining images were used for testing. To create a scenario where the training and testing images are heterogeneous, we downloaded OCT images from another study[26]. For each disease status, five images were randomly selected for creating five sets of training images. The testing set remained the same.

For brain MRI images, MRI images were downloaded from a Kaggle dataset (https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection). To create a scenario where training and testing images are homogeneous, fifteen images were randomly selected for each disease status. Five images for each disease status were used for training and the remaining images were used for testing. To create a scenario where the training and testing images are heterogeneous, we cropped either the left half or the right half of each training image. The testing set remained the same.

For chest CT images, CT images were downloaded from a Kaggle dataset (https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images). To create a scenario where training and testing images are homogeneous, fifteen images were randomly selected for each disease status. Five images for each disease status were used for training and the remaining images were used for testing. To create a scenario where the training and testing images are heterogeneous, we applied a horizontal flip to each training image. The testing set remained the same.

Image classification with LMMs

Specifics of LMMs

Image classification tasks with LMMs were all performed using the application programming interfaces (APIs) provided by the companies that developed them.

GPT-4o was performed with the `gpt-4o-2024-08-06` model via the `openai` python package (version 1.42.0) provided by OpenAI.

Claude 3.5 Sonnet was performed with the `claude-3-5-sonnet-20240620` model via the `anthropic` python package (version 0.34.2) provided by Anthropic.

Gemini 1.5 was performed with the `gemini-1.5-flash` model via the `google.generativeai` python package (version 0.8.0) provided by Google.

Image-based classification

For image-based classification, an LMM was given a reference image array consisting of two training images as the first image, the testing image as the second image, and queried with the following prompt message for tissue classification:

"There are 2 rows in total for the first image. Each row represents example images from a tissue. The second image is another image of a tissue. Which row in the first image is more similar to the second image and why?"

For cell type classification, cell state classification, retina OCT, chest CT, and brain MRI, the word "tissue" was replaced with "cell type", "cell state", "status of a tissue", "lung abnormality", and "brain abnormality" respectively in the previous prompt message.

Text-based classification

For text-based classification, GPT-4o was first given a reference image array consisting of two training images and queried with the following prompt message:

"What are the most distinct features differentiating the two rows of images? Provide detailed explanations."

The response by GPT-4o was recorded, and attached to the bottom of the following sentence to form a new prompt message.

"Is the given image more consistent with row 1 or row 2 based on the descriptions below?"

In a completely new conversation, GPT-4o was then given the testing image and queried with the above new prompt message.

Single-modal methods for image classification

The Python `torchvision` package[27] (version 0.14.1), part of the `PyTorch` project[28], was used to implement the ten single-modal methods for image classification with pre-trained weights. Specifically, the following models and their pre-trained weights were used: `alexnet` model with `AlexNet_Weights`, `densenet201` model with `DenseNet201_Weights`, `efficientnet_v2_m` model with `EfficientNet_V2_M_Weights`, `googlenet` model with `GoogLeNet_Weights`, `inception_v3` model with `Inception_V3_Weights`, `resnet50` model with `ResNet50_Weights`, `resnext101_32x8d` model with `ResNeXt101_32X8D_Weights`, `swin_b` model with `Swin_B_Weights`, `vgg16` model with `VGG16_Weights`, and `vit_b_32` model with `ViT_B_32_Weights`. The pretrained models were fine-tuned with training images provided in this study with a cross entropy loss, Adam optimizer[29], and a learning rate of 0.0001. The model training was stopped when the training loss is smaller than 0.00001 or when the training exceeds 50 epochs, whichever happens first.

Performance evaluations

Each classification result from the LMMs was manually evaluated for accuracy by a human annotator and verified by another. The classification results from the single-modal methods were evaluated programmatically.

For an image classification task, an image classification method was trained on each of the five reference image arrays and tested on each of the 20 testing images, resulting in 100 predicted results. The accuracy is defined as the percentage of the 100 predicted results that agree with the correct answers.

More formally, denote p_{ij} as the predicted binary label when a method was trained on ith ($i = 1,2,3,4,5$) reference image array and tested on jth ($j = 1,...,20$) testing image. Denote l_j as the true binary label for jth testing image. The accuracy is defined as $\sum_{i,j}(p_{ij} = l_j)/100$. The R package `ggplot2` (v.3.3.0)[30] was used for visualizing the performance.

GPT-4o word cloud

For each classification task, we calculated the number of times each word occur in the 100 responses by GPT-4o, after removing special characters. A word cloud is created using the `wordcloud` R package (version 2.6)[31] for words that occur in at least 10% of the 100 GPT-4o responses. Words without specific meanings (e.g. articles such as "a" and "the" and conjunctions such as "and") were removed manually.

Acknowledgments

Z.J. was supported by the National Institutes of Health under Award Number U54AG075936 and R35GM154865. W.H. was supported by the General Fund at Columbia University Department of Biostatistics and by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number R35GM150887. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

W.H. and Z.J. conceived the study. W.H., Y.Q., and Z.J. conducted the analysis. W.H. and Z.J. wrote the manuscript.

Competing interests

All authors declare no competing interests.

Data availability

H&E images for tissue classification were downloaded from the GTEx Histology Viewer of the GTEx portal[19] (https://gtexportal.org/home/histologyPage). Light microscopy images for cell type classification were downloaded from previous studies[20–22]. DAPI fluorescence images for cell state classification were downloaded from previous studies[23,24]. Retina OCT images for disease diagnostics were downloaded from previous studies[25,26]. Brain MRI images were downloaded from a Kaggle dataset (https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detectio
Chest CT images were downloaded from a Kaggle dataset (https://www.kaggle.com/datasets/mohamedhanyyy/

chest-ctscan-images). Responses of LMMs in this study are available as supplementary materials. All training and testing images are available at Figshare (https://figshare.com/articles/figure/Images_for_assessing_GPT-4V_s_performance_in_image_classification/24921090).

Code availability

All codes to reproduce the presented analyses are publicly available in Github repository https://github.com/Winnie09/gptimage.

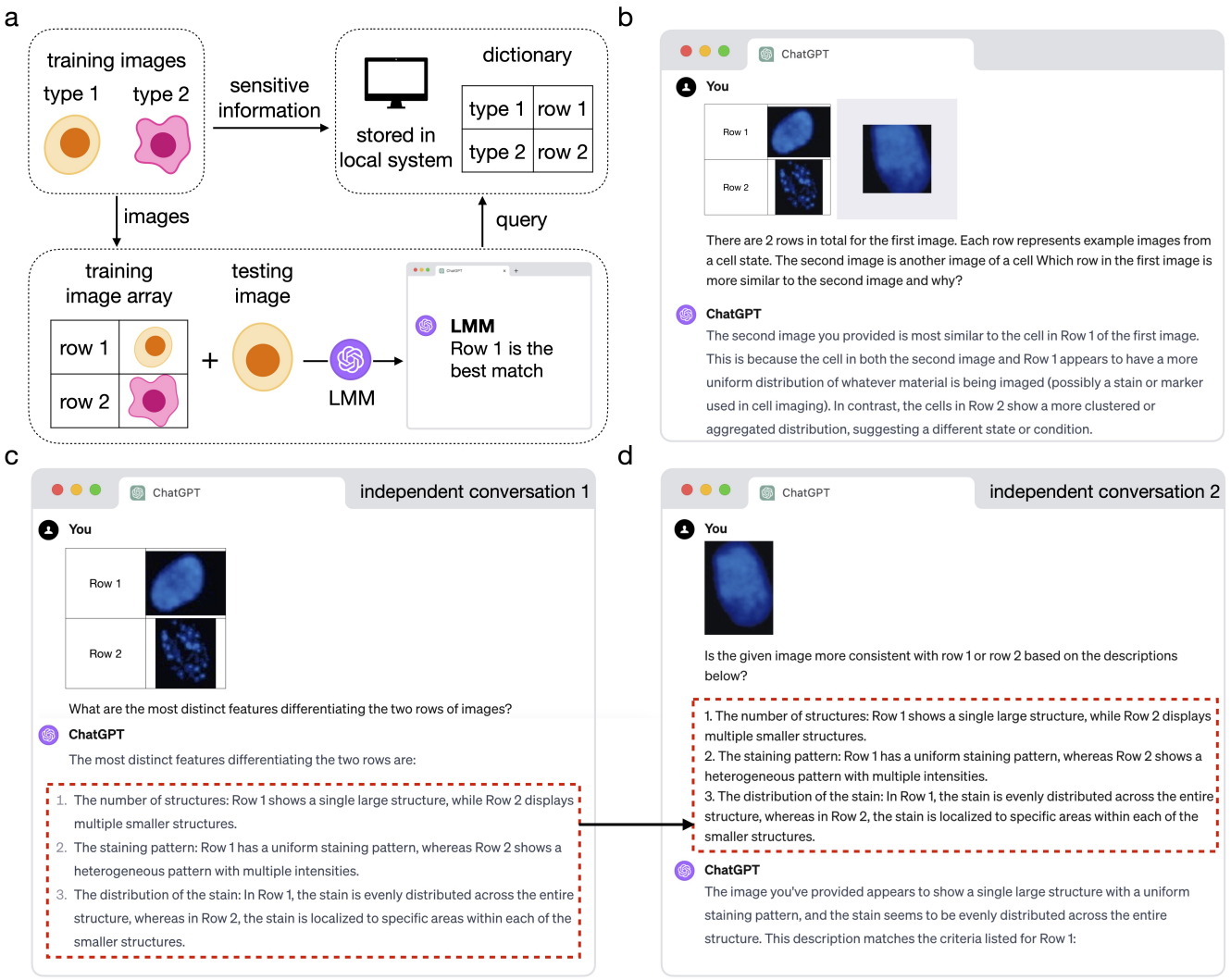


Figure 1. Demonstration of LMM image classification. **a**, A schematic of the prompt strategy for LMM image-based classification. **b**, An example conversation with GPT-4 performing image-based classification. **c, d**, An example process with GPT-4 for text-based classification. The process consists of two independent conversations: one to obtain a text description of the training image's distinct features (**c**) and another to query GPT-4 for classifying a test image (**d**).

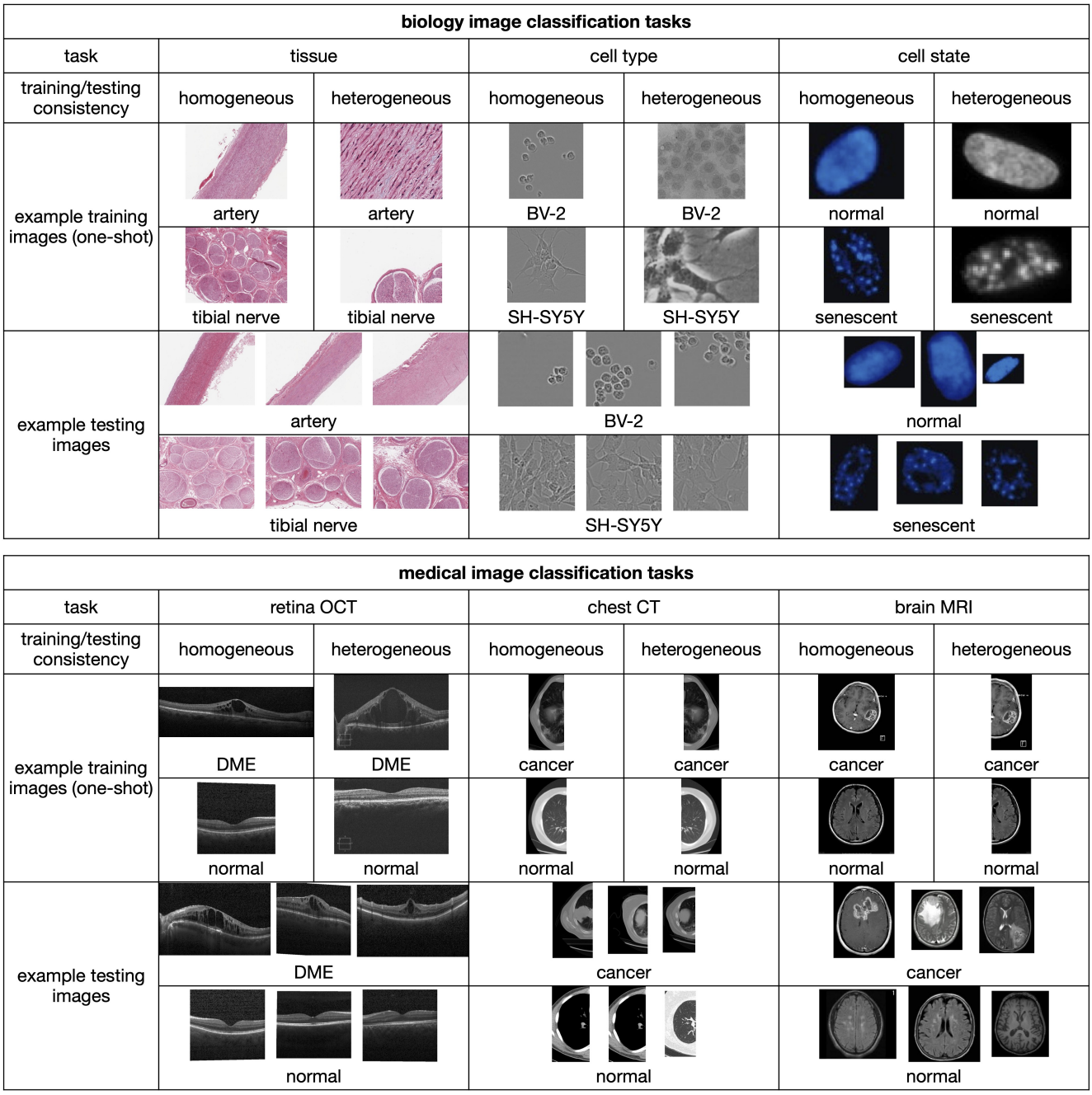


Figure 2. Example training and testing images for three biology and three medical image classification tasks.

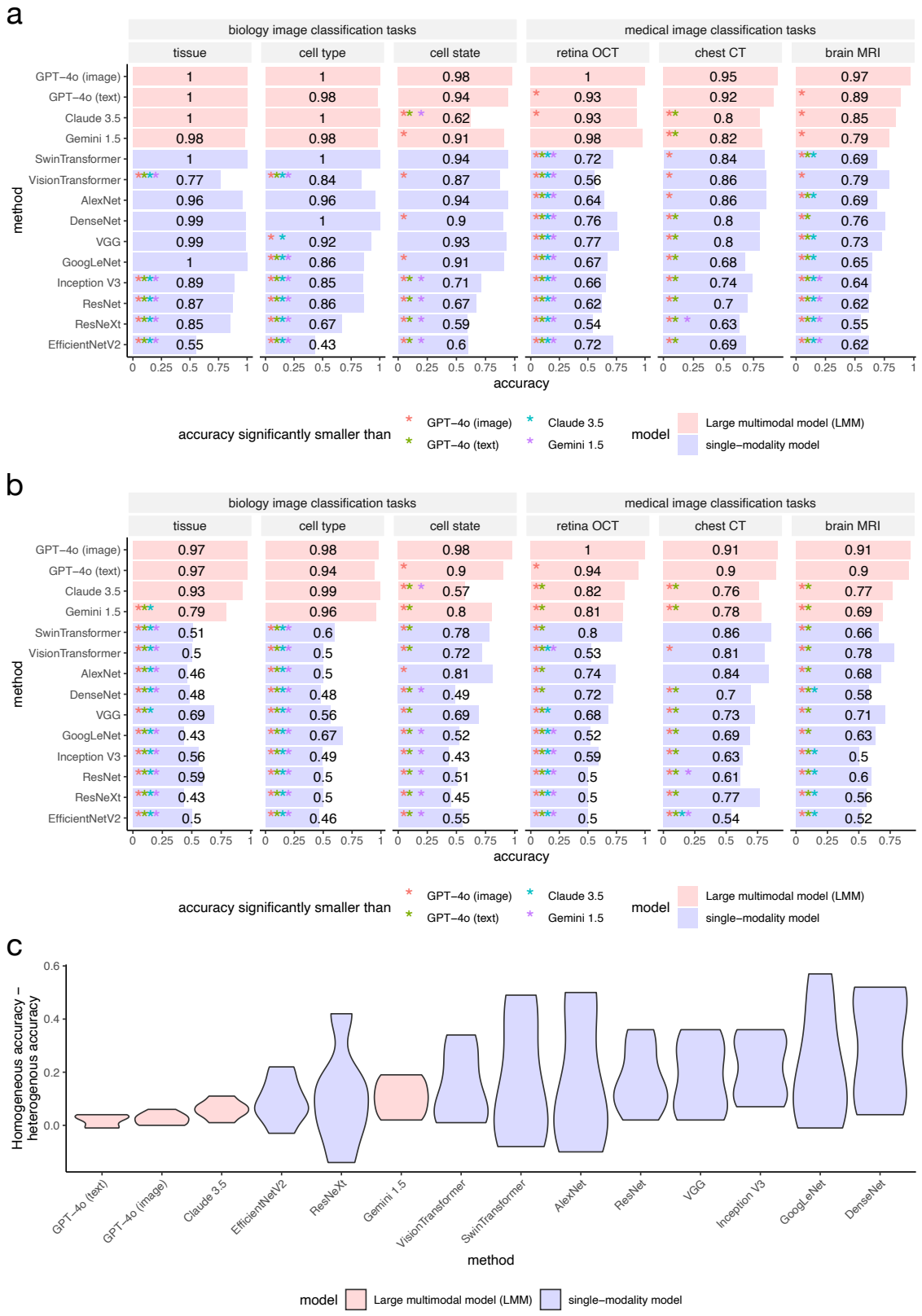
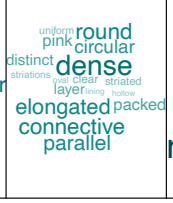
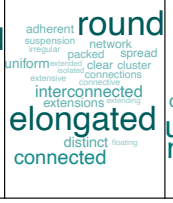


Figure 3. Performance evaluation of LMMs and single-modal methods. **a-b**, Accuracy of one-shot learning for LMMs and single-modal methods when training and testing images are homogeneous (**a**) or heterogeneous (**b**). "GPT-4o (image)" refers to image-based classification with GPT-4o, and "GPT-4o (text)" refers to text-based classification with GPT-4o. The performance of each method was compared to that of each LMM using a one-sided proportion test (R's `prop.test()` function). The p-values were adjusted using the BH procedure[32]. Methods showing significant differences are marked with an asterisk ("*"), with colors corresponding to different LMMs. **c**, Differences in accuracy when training and testing images are homogeneous or heterogeneous.

biology image classification tasks						
task	tissue		cell type		cell state	
training/testing consistency	homogeneous	heterogeneous	homogeneous	heterogeneous	homogeneous	heterogeneous
word cloud						

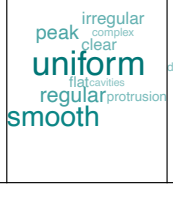
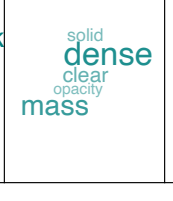
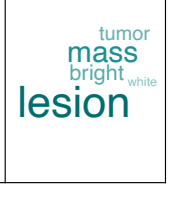
medical image classification tasks						
task	retina OCT		chest CT		brain MRI	
training/testing consistency	homogeneous	heterogeneous	homogeneous	heterogeneous	homogeneous	heterogeneous
word cloud						

Figure 4. Word cloud showing keywords from GPT-4o's responses in image-based classification.

References

1. Achiam, J. *et al.* Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
2. Reid, M. *et al.* Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).
3. Yang, Z. *et al.* The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421* **9**, 1 (2023).
4. Yin, S. *et al.* A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549* (2023).
5. Wu, J., Gan, W., Chen, Z., Wan, S. & Philip, S. Y. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, 2247–2256 (IEEE, 2023).
6. Liu, Z. *et al.* Holistic evaluation of gpt-4v for biomedical imaging. *arXiv preprint arXiv:2312.05256* (2023).
7. Liu, Y. *et al.* A systematic evaluation of gpt-4v's multimodal capability for chest x-ray image analysis. *Meta-Radiology* 100099 (2024).
8. Yan, Z. *et al.* Multimodal chatgpt for medical applications: an experimental study of gpt-4v. *arXiv preprint arXiv:2310.19061* (2023).
9. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. neural information processing systems* **25** (2012).
10. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708 (2017).
11. Tan, M. & Le, Q. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, 10096–10106 (PMLR, 2021).
12. Szegedy, C. *et al.* Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9 (2015).
13. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826 (2016).
14. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
15. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500 (2017).
16. Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022 (2021).
17. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
18. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
19. GTEx Consortium. The Genotype-Tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
20. Edlund, C. *et al.* LIVECell-A large-scale dataset for label-free live cell segmentation. *Nat. Methods* **18**, 1038–1045 (2021).
21. Saleppico, S. *et al.* Iron regulates microglial cell-mediated secretory and effector functions. *Cell. Immunol.* **170**, 251–259 (1996).
22. Ross, R. A., Spengler, B. A. & Biedler, J. L. Coordinate morphological and biochemical interconversion of human neuroblastoma cells. *J. Natl. Cancer Inst.* **71**, 741–747 (1983).
23. Narita, M. *et al.* Rb-mediated heterochromatin formation and silencing of E2F target genes during cellular senescence. *Cell* **113**, 703–716 (2003).
24. Nelson, D. M. *et al.* Mapping H4K20me3 onto the chromatin landscape of senescent cells indicates a function in control of cell senescence and tumor suppression through preservation of genetic and epigenetic stability. *Genome Biol.* **17**, 158 (2016).
25. Kermany, D. S. *et al.* Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131.e9 (2018).

26. Gholami, P., Roy, P., Parthasarathy, M. K. & Lakshminarayanan, V. Octid: Optical coherence tomography image database. *Comput. Electr. Eng.* **81**, 106532 (2020).

27. Marcel, S. & Rodriguez, Y. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM international conference on Multimedia*, 1485–1488 (2010).

28. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. neural information processing systems* **32** (2019).

29. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

30. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. Use R! (Springer International Publishing, Switzerland, 2016).

31. Oesper, L., Merico, D., Isserlin, R. & Bader, G. D. Wordcloud: a cytoscape plugin to create a visual semantic summary of networks. *Source code for biology medicine* **6**, 7 (2011).

32. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).