npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01893-8

Specialized curricula for training vision language models in retinal image analysis

Robbie Holland[1] ✉, Thomas R. P. Taylor[2], Christopher Holmes[3], Sophie Riedl[4,5], Julia Mai[4,5], Maria Patsiamanidi[2], Dimitra Mitsopoulou[2], Paul Hager[5], Philip Müller[5], Johannes C. Paetzold[1,6], Hendrik P. N. Scholl[7,8,9], Hrvoje Bogunović[10], Ursula Schmidt-Erfurth[4], Daniel Rueckert[1,6,11,14], Sobha Sivaprasad[3,14], Andrew J. Lotery[2,14] & Martin J. Menten[1,6,11,14] On behalf of the PINNACLE consortium*

Clinicians spend significant time reviewing medical images and transcribing findings. By integrating visual and textual data, foundation models have the potential to reduce workloads and boost efficiency, yet their practical clinical value remains uncertain. In this study, we find that OpenAI's ChatGPT-4o and two medical vision-language models (VLMs) significantly underperform ophthalmologists in key tasks for age-related macular degeneration (AMD). To address this, we developed a dedicated training curriculum, designed by domain specialists, to optimize VLMs for tasks related to clinical decision making. The resulting model, RetinaVLM-Specialist, significantly outperforms foundation medical VLMs and ChatGPT-4o in AMD disease staging (F1: 0.63 vs. 0.33) and referral (0.67 vs. 0.50), achieving performance comparable to junior ophthalmologists. In a reader study, two senior ophthalmologists confirmed that RetinaVLM's reports were substantially more accurate than those written by ChatGPT-4o (64.3% vs. 14.3%). Overall, our curriculum-based approach offers a blueprint for adapting foundation models to real-world medical applications.

Medical images are central to many clinical decisions regarding patient diagnosis, referral, and treatment. Clinicians spend a significant amount of time transcribing image-based decisions into text in order to store and communicate their findings[1,2]. Visual-language models (VLM), which automatically interpret medical images and generate detailed textual descriptions, have enormous potential to alleviate clinical workloads and increase patient access to high-quality medical care[3,4]. To date, the majority of medical VLMs have been trained to output a finite set of pre-determined textual responses[5–8]. Only recently, the combination of large language models (LLM) with medical vision encoders has led to the development of more powerful and versatile *generative* VLMs that are able to write comprehensive text reports or answer complex questions[9–11].

This current generation of medical language models is fueled by vast amounts of unstructured training data that is extracted from medical textbooks, scientific publications or social media posts of healthcare professionals[6,9,10]. These *foundation* language models have stirred considerable interest among the medical community for their expert-level performance on standardized medical question-answering tasks, such as licensing exams and case studies[12,13]. However, it is unclear whether this general performance translates to clinical utility in specialist medical domains[14]. Despite its impressive scale, the training data of foundation language models has been collected agnostically of their downstream application. The resulting model is unlikely to acquire the nuanced knowledge necessary for effective application in specialized clinical contexts.

In this study, we identify this missing piece in foundation models towards developing generative medical VLMs with real-world clinical utility. We propose to deconstruct clinical problems into sets of mandatory capabilities required for their resolution and selectively train VLMs in these

[1]Biomedical Image Analysis, Department of Computing, Imperial College London, London, United Kingdom. [2]Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Southampton, United Kingdom. [3]Moorfields Eye Hospital NHS Foundation Trust, London, United Kingdom. [4]Ophthalmic Image Analysis Group (OPTIMA), Medical University of Vienna, Vienna, Austria. [5]Department of Ophthalmology and Optometry, Medical University of Vienna, Vienna, Austria. [6]Chair for AI in Healthcare and Medicine, Technical University of Munich, Munich, Germany. [7]Institute of Molecular and Clinical Ophthalmology Basel, Basel, Switzerland. [8]Department of Ophthalmology, University of Basel, Basel, Switzerland. [9]Department of Clinical Pharmacology, Medical University of Vienna, Vienna, Austria. [10]Institute of Artificial Intelligence, Centre for Medical Data Science, Medical University of Vienna, Vienna, Austria. [11]Munich Center for Machine Learning, Munich, Germany. [14]These authors contributed equally: Daniel Rueckert, Sobha Sivaprasad, Andrew J. Lotery, Martin J. Menten. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: robbie.holland@stanford.edu

skills. To train VLMs in these skills, we develop a curriculum-based approach[15] that draws from recent advances in instruction finetuning[16–18] which iteratively refine models on datasets of increasing quality. We demonstrate the feasibility of this approach in ophthalmology, focusing our analysis on a single retinal disease in order to assess the clinical limitations of foundation VLMs, and the benefits of training on specialized curricula, in depth. To this end we introduce, RetinaVLM, is a generative medical VLM for OCT images (see Fig. 1a). RetinaVLM is trained using a two-part dedicated curriculum that is specific to the clinical management of age-related macular degeneration (AMD), the leading cause of blindness in the elderly[19,20]. The resulting model is able to process optical coherence tomography (OCT) images of the retina and flexibly respond to instructions and questions (see Fig. 1b). In particular, we evaluate RetinaVLM's utility and versatility regarding disease staging, patient referral and biomarker analysis in AMD (see Fig. 1c).

Results

RetinaVLM, a specialist vision-language model for retinal image analysis

RetinaVLM combines two main components: an ophthalmic vision encoder that processes input OCT images, and a generative LLM that handles textual instructions and outputs the corresponding responses (see Fig. 1a). The vision encoder was trained using self-supervised learning on images from the train set, and was found to perform on par with RETFound[21], a large foundation model for retinal image analysis[22]. For the language model, we use Meta's Llama 3 as generative LLM which was the most performant, openly available model at the time of this study[23]. Both these deep neural networks have already been pre-trained on large OCT and natural language datasets, respectively. We combine these to create RetinaVLM, following the architectural design of MiniGPT-4 introduced by Zhu et al.[24]. This approach leaves the pretrained models unchanged, and instead trains a third, intermediary network to map visual information from the image encoder to the language model. For additional information regarding the model architecture, training and inference see the "RetinaVLM vision-language model architecture" section.

A curriculum to encode the capabilities of retinal specialists into vision-language models

An intuitive strategy to specialize VLMs while preserving their ability to flexibly interact with text queries is to provide them with a set of medical images and corresponding *visual question-answer* (VQA) pairs. VQA-based training, a form of instruction fine-tuning[18,25], guides VLMs to extract visual information and synthesize it into textual outputs that adhere to specific user directives. This approach also introduces diversity through varying instructions and responses, mitigating overfitting while enhancing the model's capacity to selectively report relevant clinical features and contextualize findings. However, relevant VQA datasets are scarce for most medical specializations, including ophthalmology.

Together with a large team of ophthalmologists, which are involved with the patient care and academic research of AMD, we created a curriculum of VQA datasets designed to train VLMs for assisting image-based clinical decisions regarding AMD. To this end, the ophthalmologists first defined a set of guidelines outlining essential capabilities of agents assisting the image-based clinical management of AMD (verbose versions are documented in Supplementary Fig. 10). Specifically, these include details relating to the identification of AMD biomarkers in OCT images, the linking of these to the AMD disease stage, and ultimately deciding on the required referral and treatment of the patient. These subsequently guided a combination of manual and automated efforts to curate a training curriculum, which consists of 41,926 OCT images, and 479,710 VQA pairs to progressively specialize VLMs in these capabilities.

The first part of the curriculum, named *Introduction to retina*, primarily covers the appearance of the retina and AMD biomarkers in OCT images. Using automated data collection, we obtained tabular reports for 41,926 retrospectively collected OCT images of AMD patients (see Fig. 2a). Each report describes the visible biomarkers, patient's diagnosis, visual acuity and demographic information in 34 data fields. A full description of the retrospective OCT dataset and tabular field collection process can be found in the "Retrospective patient dataset" section. Four example tabular reports are shown in Supplementary Fig. 1.

Next, we tasked an independent LLM to generate question-answer pairs based on these reports (see Fig. 2c). The model processed the content of the tabular reports – but not the OCT images – to output a numbered list of question-answer pairs. We generated an average of ten question-answer pairs per report that are mostly related to the presence or absence of specific biomarkers (see Fig. 2e). The LLM was instructed to create both closed-ended 'yes or no' style questions, and simple open-ended questions. Detailed information on the LLM setup can be found in the "Report curation and question-answer pair generation (curriculum part 1)" section.

This automated approach allowed us to generate a large dataset of 408,545 question-answer pairs. However, the questions were limited in scope to the set of biomarkers documented by the tabular reports. Training on these yielded the first of two specialist VLMs, *RetinaVLM-Base* (see Fig. 2g).

The second part of the curriculum, named *Advanced retinal specialism*, builds on top of the first part to link imaging biomarkers to AMD stage and the recommended course of treatment. As this reasoning cannot be fully conveyed via tabular information, we tasked two ophthalmologists with 3 and 10 years of experience, respectively, to create comprehensive textual reports for a subset of 330 OCT images (see Fig. 2b). The ophthalmologists were asked to primarily describe the main pathological biomarkers related to AMD while also noting any other observations regarding the retinal anatomy. This task yielded high-quality reports that go beyond the short notes that are typically written by ophthalmologists in their clinical routine. Instructions given to the ophthalmologists are provided in the "Report curation and question-answer pair generation (curriculum part 2)" section, and six sample reports yielded by this process are shown Supplementary Fig. 2.

Similar to before, an independent LLM was then employed to automatically generate question-answer pairs based on the reports (see Fig. 2d). Due to the substantially increased depth and scope of the full-text reports compared to the tabular ones, we used several advanced LLM instructions to create 216 diverse question-answer pairs per image on average (see Fig. 2f). These cover additional biomarkers and sub-categorize them based on their size, type, and location. Other question-answer pairs are related to the causal relationship between biomarkers and six AMD disease stages as well as three levels of patient referral urgency. Moreover, the question-answer pairs were more varied in their structure in order to preserve interactive capabilities of the foundation LLM. For example, some queries asked to summarize the existing reports or provide several answers in succession. An example interaction with the LLM to generate question-answer pairs with the LLM is shown in Supplementary Fig. 3. Furthermore, a list of all the LLM instructions is provided in "Question-answer generation prompts" in the Supplementary material and example question-answers yielded by this approach are shown in the 'part 2' section of Supplementary Fig. 4.

This resulted in a dataset of 71,165 advanced question-answer pairs. By further training RetinaVLM-Base on the second part of the curriculum, we obtained our most performant VLM for the clinical management of AMD, *RetinaVLM-Specialist* (see Fig. 2h).

RetinaVLM-Specialist outperforms foundation models and approaches junior ophthalmologists in AMD disease staging and report writing

AMD is a debilitating and irreversible condition marked by the progressive loss of central vision, severely hindering essential activities like reading, driving, and recognizing faces. The demand for timely diagnosis and management currently exceeds the available ophthalmology expertise[26]. Moreover, largely due to aging populations, projections indicate a nearly 50% increase in AMD cases globally to nearly 300 million by 2040[20]. In this context, automated retinal image analysis emerges as an

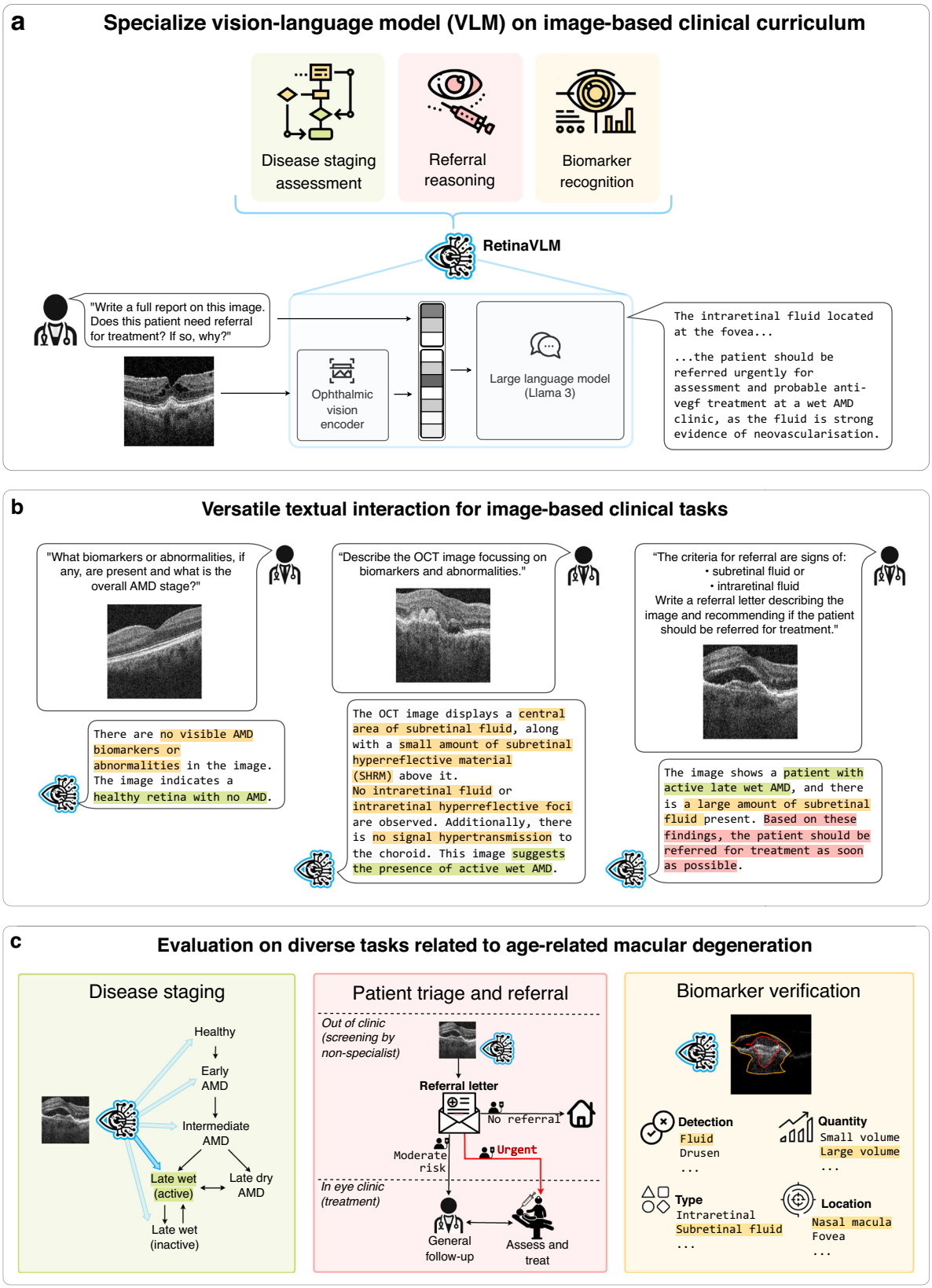


Fig. 1 | A curriculum-based approach for training vision-language models in medical specialties. We introduce RetinaVLM, a specialist medical generative vision-language model (VLM). **a** Using a curriculum-based approach, we trained RetinaVLM in specialist medical skills that medical foundation VLMs are currently lacking. **b** RetinaVLM is able to process retinal optical retinal optical coherence tomography (OCT) images and flexibly respond to text-based queries. **c** Its abilities entail the analysis of imaging biomarkers of age-related macular degeneration (AMD), disease staging, and the referral for treatment.

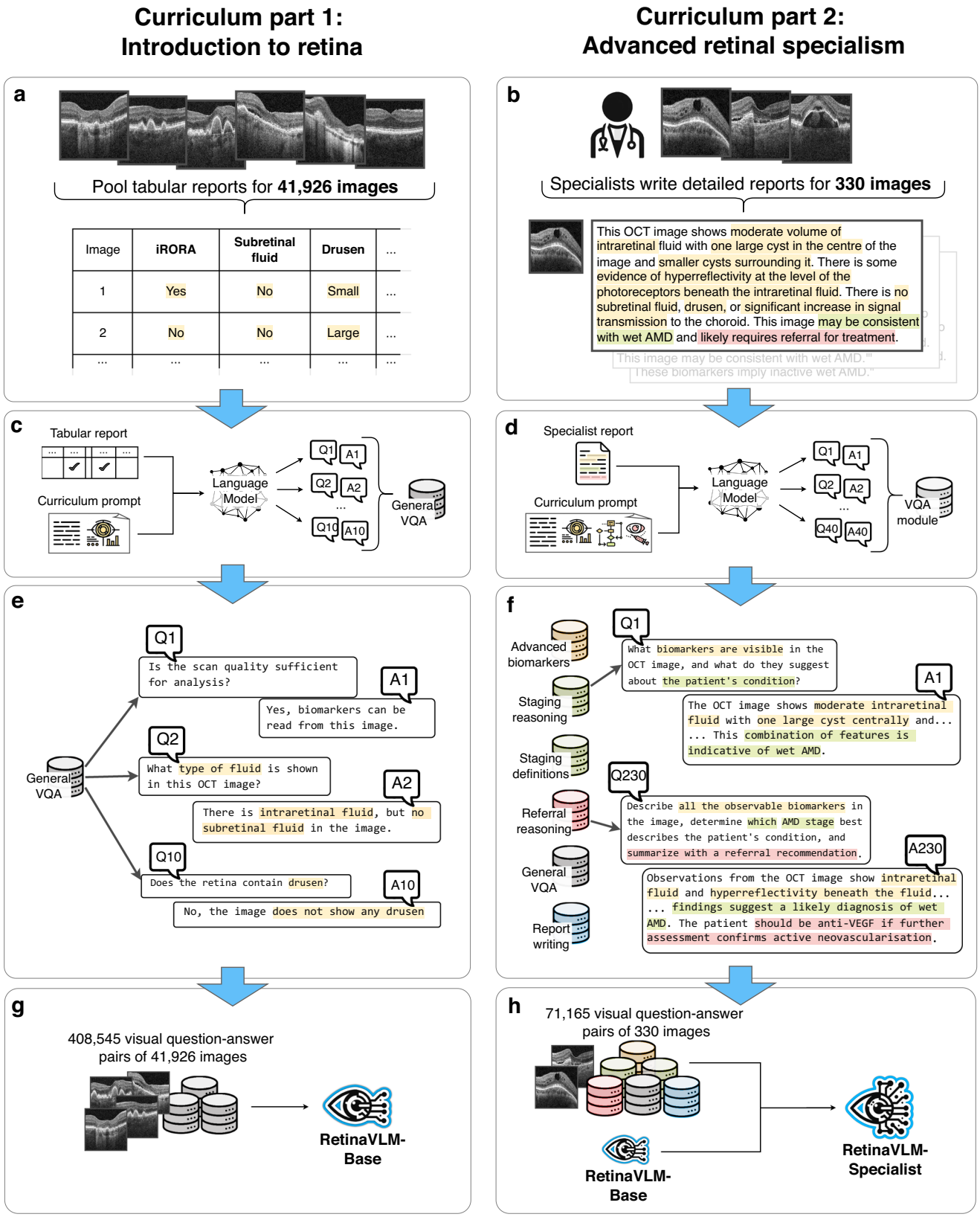


Fig. 2 | Creating medical visual question-answers for a two-part training curriculum. We curated a two-part curriculum to specialize medical VLMs for clinical use. **a**, **b** Based on a retrospectively collected OCT imaging dataset, we created a large number of tabular reports as well as a small number of comprehensive textual reports. **c**, **d** We then used an independent LLM to automatically generate visual question-answers based on these reports. **e**, **f** This yielded two VQA datasets, the first on basic imaging biomarkers of AMD and the second covering more advanced clinical skills. **g**, **h** Finally, we trained two specialist medical generative VLMs, RetinaVLM-Base and RetinaVLM-Specialist, using either the first or both VQA datasets.

essential tool to support the interpretation and textual reporting of retinal images.

A key aspect of image report generation involves estimating the disease stage indicated by the retinal image. We assessed the ability of five different generative VLMs to determine the AMD disease stage when writing reports on retinal OCT images. Specifically, we benchmarked two foundation VLMs, Med-Flamingo[10] and LLaVA-Med[9], with purported general abilities in medical image analysis. In addition, we assessed OpenAI's ChatGPT-4o model[27], developed by OpenAI, L.P. (San Francisco, California, USA). We compared these baseline models against RetinaVLM-Base and RetinaVLM-Specialist, which result from cumulative training on curriculum part 1 and part 2, respectively. Finally we compared these automated models against the overall performance of the six junior ophthalmologists. This experiment was conducted using a testing dataset of 276 previously unseen OCT images, on which VLMs were tasked to write descriptive reports before classifying the patient into one of six disease stages (see Fig. 3a). The model predictions were compared to ground truth labels obtained from ophthalmologists. Each image was initially graded by two out of six junior ophthalmologists, who have 2, 3, 5, 8, 10 and 15 years of experience working full-time in ophthalmology clinics after receiving their medical degree. Inter-rater disagreements were resolved by a panel of two senior ophthalmologists with 25 and 32 years of experience (see "Definitions and roles of the junior and senior ophthalmologists" in the Supplementary material). Further details regarding the derivation of testing labels are provided in the "Retrospective patient dataset" section, and the instruction given to all VLMs in the "Report generation for disease staging" section.

We found that both the intermediate RetinaVLM-Base model and ChatGPT-4o perform significantly better than Med-Flamingo and LLaVA-Med, which lack the ophthalmological specialism to stage disease (see Fig. 3b). By more effectively classifying conversion to late stages of AMD RetinaVLM-Base and ChatGPT-4o achieve F1 scores of 0.33 and 0.29, respectively. However, beyond this distinction, these models failed to differentiate finer disease stage variations and were markedly outperformed by RetinaVLM-Specialist which scored 0.63 F1. This performance approached, but did not match, the accuracy of the junior ophthalmologists who achieved an F1 score of 0.78. We analyze this last discrepancy in further detail in the Discussion. Moreover, foundation VLMs and RetinaVLM-Base returned a substantial number of invalid reports that did not conclude with one of the six disease stages (see Fig. 3c). Conversely, all generated reports by RetinaVLM-Specialist were valid. Similar to human experts, RetinaVLM-Specialist struggled the most when diagnosing wet inactive AMD. We attribute this to the higher number of imaging biomarkers shared by intermediate, inactive late wet, and active late wet AMD, which sometimes led to misinterpretation by both ophthalmologists and RetinaVLM-Specialist (see Fig. 3c, d). Four representative examples of success and failure cases of RetinaVLM-Specialist are shown in Supplementary Fig. 5a. Moreover, full numerical results, including a comparison to a standard image-only classification model, as well as the confusion matrix for medical foundation models, are shown in Supplementary Figs. 6a and 7, respectively.

Next, we performed a qualitative evaluation of imaging reports by senior ophthalmologists. Eight-four of the generated reports were scored by the two senior ophthalmologists for their correctness, completeness, and conciseness. They were shown 28 reports written by ChatGPT-4o, 28 by RetinaVLM-Specialist, and 28 by the two annotating junior ophthalmologists. We then randomly ordered the 84 reports to decrease the likelihood that multiple reports regarding the same image would appear in succession. Moreover, to mitigate bias we conducted this evaluation as a single-blind study in which the authorship of each report was concealed from the senior ophthalmologists. For each report, the senior ophthalmologist first reviewed the corresponding OCT image before rating the generated report in the three criteria on a five point Likert scale[28]. For details on the report generation instruction and evaluation criteria see the "Report generation for qualitative evaluation" section

The senior ophthalmologists observed that ChatGPT-4o largely failed to compose factually correct image reports (see Fig. 4a), even though it uses specialist terminology that may give the reports the initial appearance of being written by an ophthalmologist (see Fig. 4b). ChatGPT-4o was found to consistently hallucinate the presence of subretinal fluid (further instances of hallucinations are shown in Supplementary Fig. 13). Moreover, in every relevant instance ChatGPT-4o failed to detect presence of subretinal hyperreflective material and hypertransmission, which are crucial for diagnosing inactive late wet and late dry AMD, respectively. Further examples of errors made by ChatGPT-4o can be found in in Supplementary Fig. 14, including verbose versions of the three reports displayed in Fig. 4b. Overall, the senior ophthalmologists found that only 4 out of 28 (14.3%) of the reports written by ChatGPT-4o were correct in their observations and conclusions.

Overall senior ophthalmologists either agreed or strongly agreed that reports generated by RetinaVLM-Specialist were more correct (18 vs. 4, or 64.3% vs. 14.3%), complete (18 vs. 5) and concise (18 vs. 4) than those written by ChatGPT-4o. Compared to ChatGPT-4o, RetinaVLM-Specialist correctly detected a wider range of biomarkers, which more frequently led to a correct disease stage estimation. However, there remains a gap in overall performance between RetinaVLM-Specialist and the junior ophthalmologists, with the senior ophthalmologists rating their reports to be more correct (25 vs. 18), complete (24 vs. 18) and concise (22 vs. 18).

An example of this discrepancy can be seen in the second sample in Fig. 4b, where RetinaVLM-Specialist correctly identified that the image showed a healthy retina, but also detects a small cyst that was not found in the image (for additional examples of hallucinations see Supplementary Fig. 13). Junior ophthalmologists wrote a concise report, but incorrectly associated subretinal drusenoid deposits with intermediate AMD. Moreover, both RetinaVLM, which has currently only been trained to identify AMD from retinal imaging biomarkers, and ChatGPT-4o fail to report likely cases of retinal vascular disease. These characteristics of both the junior ophthalmologists and RetinaVLM-Specialist are discussed in more detail in the Discussion.

RetinaVLM-Specialist surpasses opticians and approaches junior ophthalmologists in AMD patient screening and referral

As the prevalence of AMD is expected to further increase in the upcoming decades[20], ocular screening programs are being introduced around the world. In the United Kingdom, some projects involve opticians and pharmacies that acquire and interpret OCT images. They may refer a patient to a specialist clinic, summarizing their findings and the estimated level of the patient's risk in a letter. In the United Kingdom, treatment guidelines for AMD mandate that patients with signs of neovascularization are referred for immediate treatment within two weeks. However, non-specialists exhibit a tendency to over-diagnose these cases. An internal audit at Southampton Eye Unit found that 74.2% of the referrals made to the clinic do not have any form of treatable AMD. The processing and assessment of these false positives affects the clinic's ability to care for the remaining patients with treatable forms of AMD.

We evaluated the ability of VLMs to assess the level of referral urgency from OCT image (see Fig. 5a). For each case, the VLMs were provided explicit referral guidelines, and asked to recommend which of three levels of referral urgency was most appropriate for the patient: *no referral* for healthy patients, *to be seen within 18 weeks (routine referral)* for patients that are at risk of progressing to active late wet AMD but do not require treatment yet, and *referral within two weeks* for patients with any signs of neovascularization that should be urgently referred for antiangiogenic treatment. Two junior ophthalmologists independently reviewed images of 95 patients that have previously been referred to the hospital for treatment of wet AMD. For each patient, they independently decided the most appropriate of the three levels of referral urgency, and disagreements were arbitrated by the two senior ophthalmologists. In line with previous audits, they found the false discovery rate for urgent referrals was 69.5%. We then calculated F1 scores for the highest risk patients in need of urgent referral between the VLM's predictions and the ground truth. The full referral protocol and report generation instruction are provided in the "Report generation for patient referral" section.

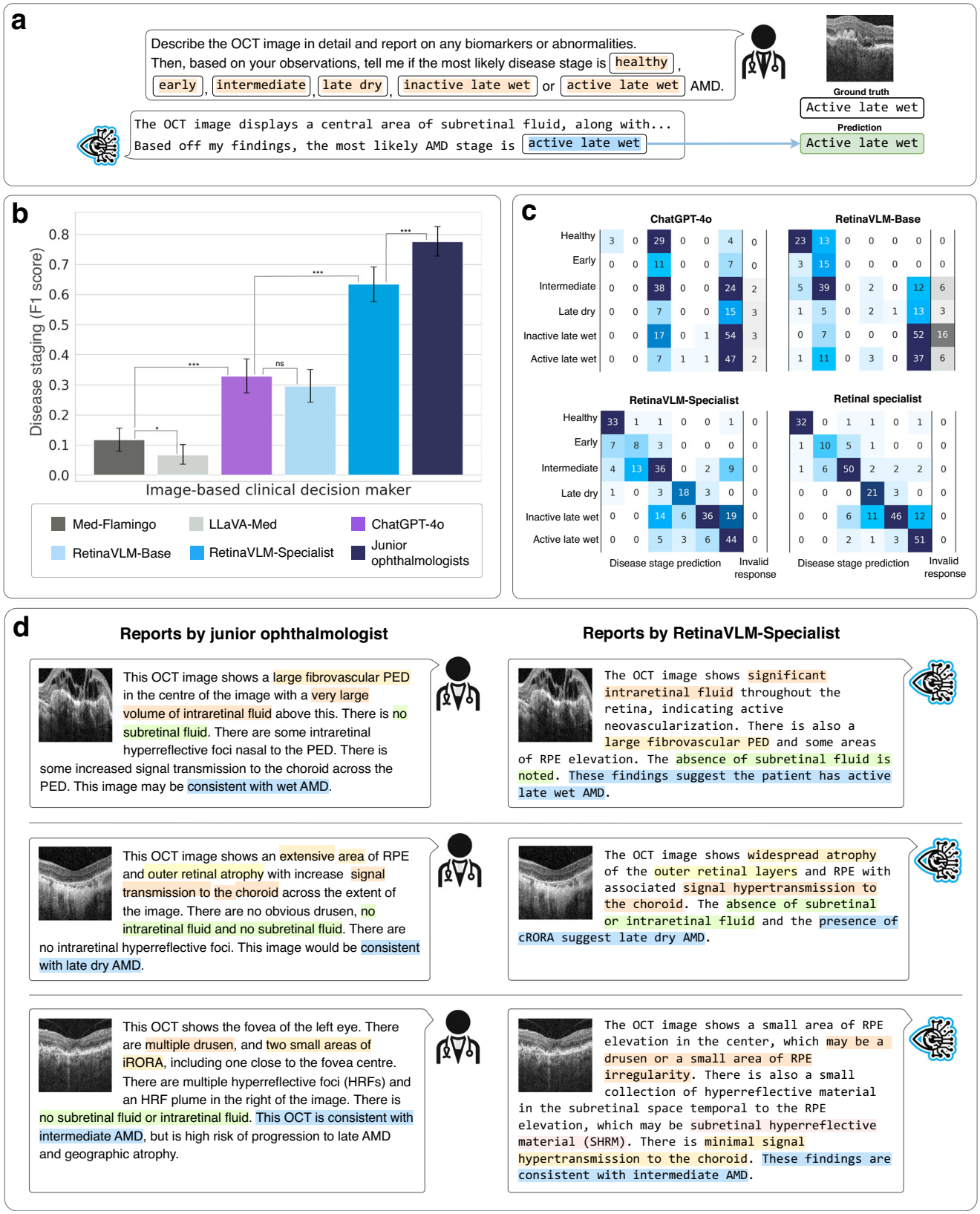
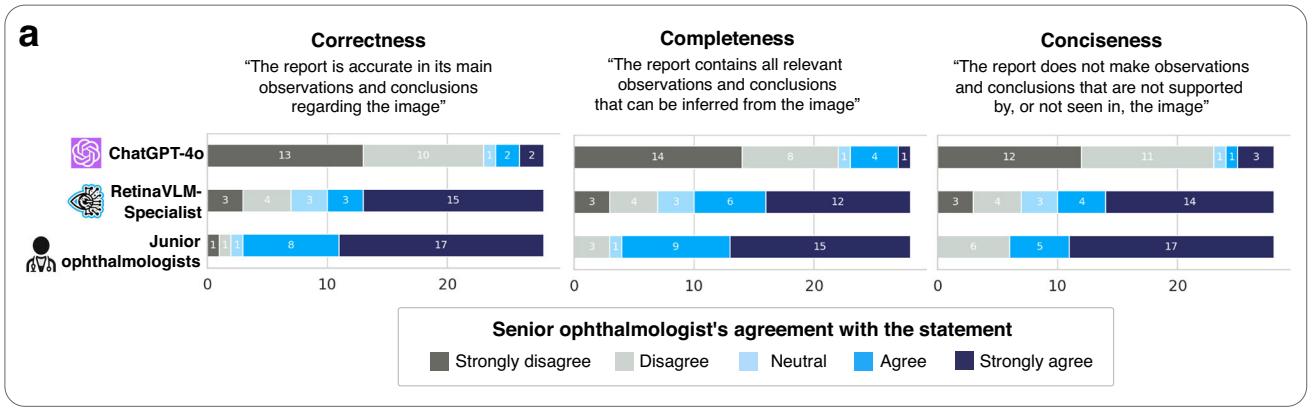


Fig. 3 | Comparison of vision-language models with ophthalmologists in reporting AMD disease stage. a Comparison of the ability of four VLMs to write reports on retinal OCT images and derive the AMD stage. **b** Overall staging accuracy for each model was calculated using micro F1 scores with 95% CI, with tests of statistical significance calculated using McNemar's test. **c** Confusion matrices between the senior ophthalmologists' assessments (rows) against the image-based clinical decision maker's prediction (columns). **d** Qualitative comparison of reports written by human ophthalmologists and RetinaVLM-Specialist with text markings highlighting findings regarding biomarker observations and disease stage.



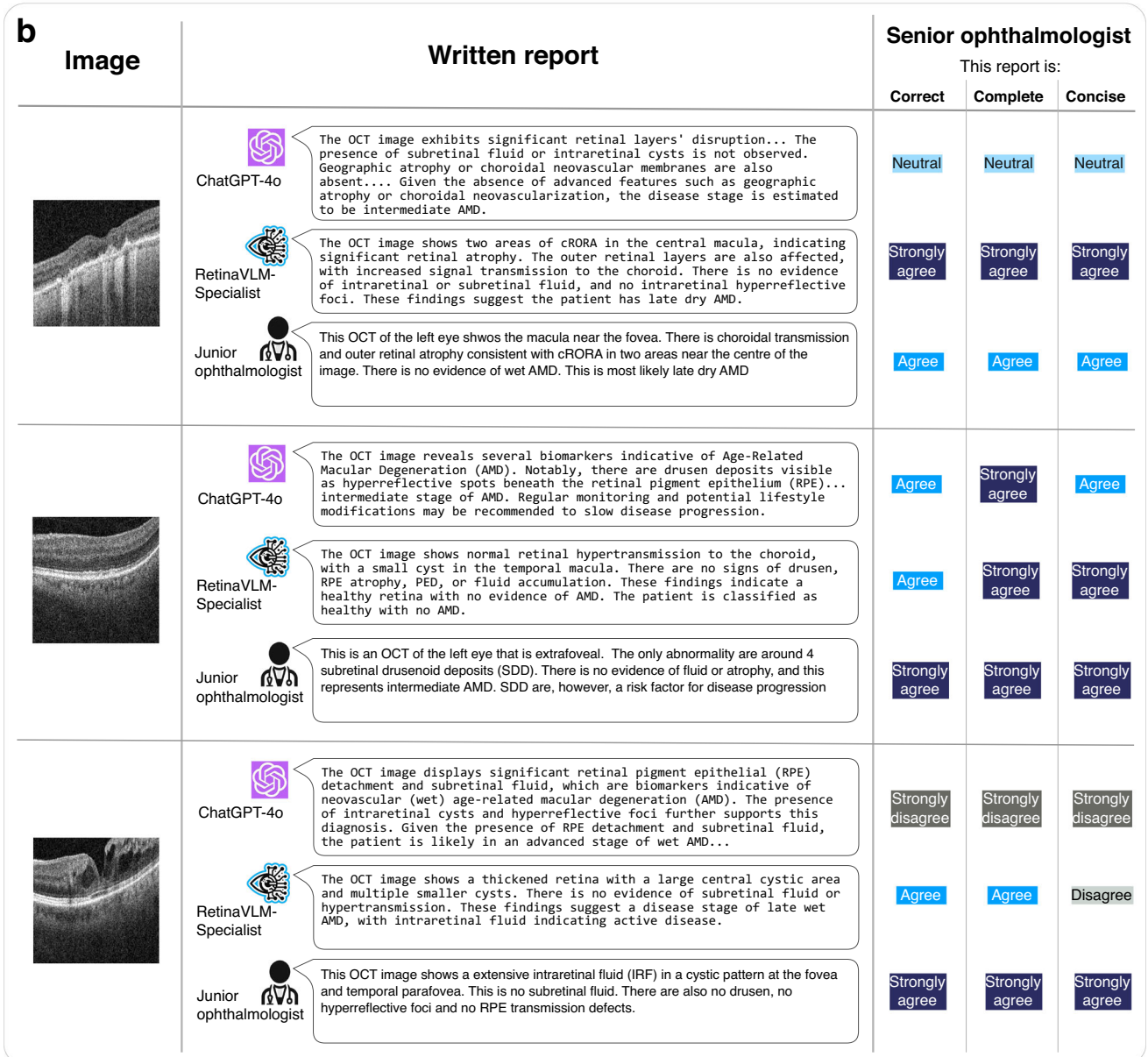


Fig. 4 | Qualitative evaluations of image reports written by vision-language models and ophthalmologists. a Summary statistics of the quality of image reports written by ChatGPT-4o, RetinaVLM-Specialist and junior ophthalmologists, broken down by correctness, completeness and conciseness. Reports were scored for on each of the three criteria by senior ophthalmologists using a five-point Likert scale. **b** Representative reports with ratings by one of the senior ophthalmologists. As ChatGPT-4o tended to write excessively long reports, despite being prompted to shorten them, we display passages the senior ophthalmologists selected as the most important to their given rating. For verbose versions and additional sample reports by ChatGPT-4o see Supplementary Fig. 14.

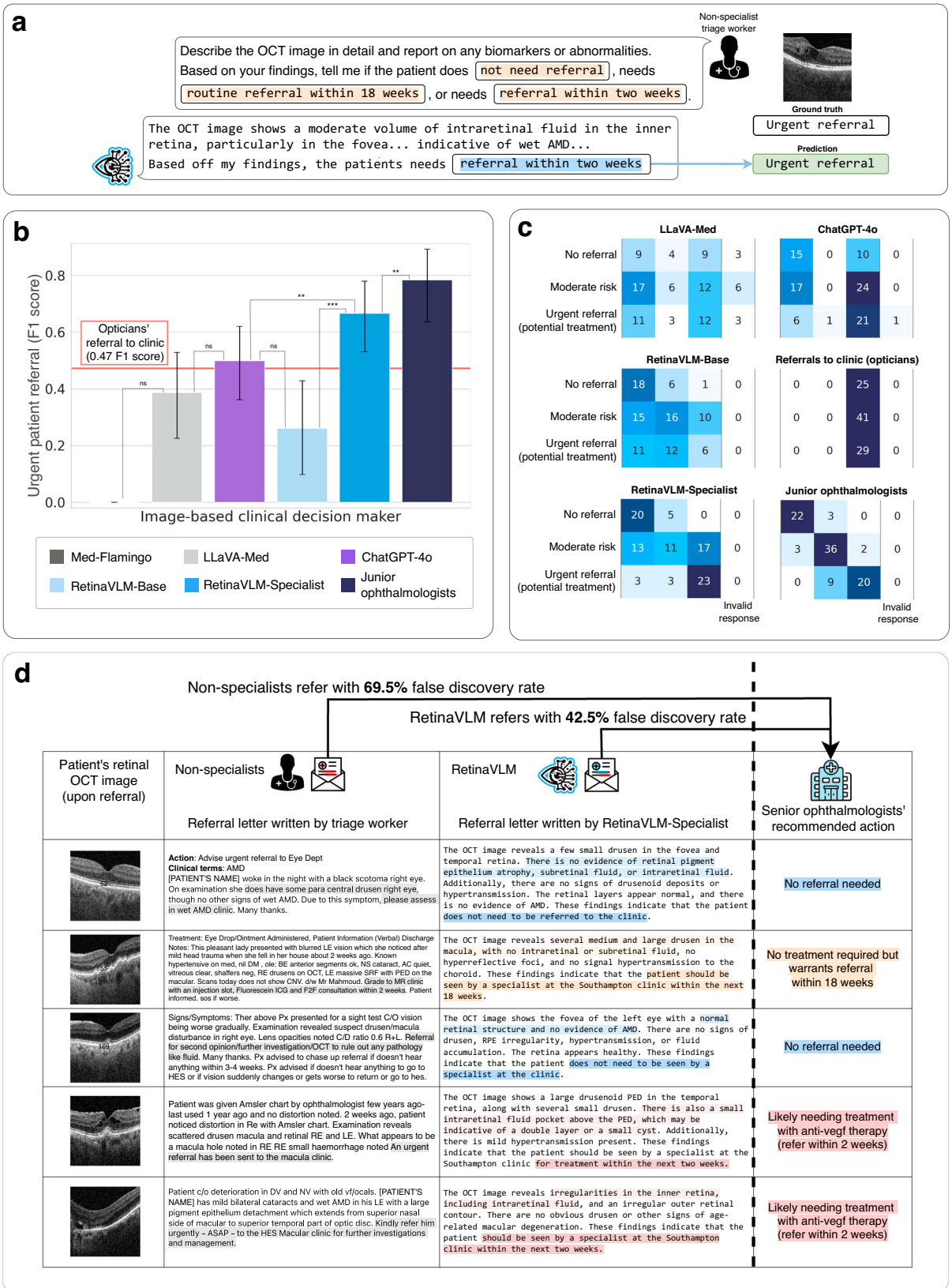


Fig. 5 | Comparison of vision-language models with ophthalmologists in writing image-based referral reports. a Evaluation of the ability of four VLMs to assess the need for patient referral for treatment of wet AMD. **b** Overall referral accuracy was calculated using F1 score for urgent referral with a 95% CI. Tests of statistical significance were carried out using McNemar's test. **c** Confusion matrices between the senior ophthalmologists assessment (rows) against the image-based clinical decision maker's referral assessment (columns). **d** Image reports written by the non-specialist optician who originally referred the patient, compared with reports of the same patient written by RetinaVLM-Specialist.

We found that both medical foundation VLMs and Retina-Base perform worse than opticians regarding their ability to refer patients in need of urgent treatment (see Fig. 5b). While Med-Flamingo failed to refer any of the 29 high-risk patients cases, LLaVA-Med and RetinaVLM-Base were ineffective for differentiating high-risk patients from low- to moderate-risk patients (see Fig. 5c). In comparison, ChatGPT-4o was relatively more effective for detecting urgent referrals, but still recommended the referral of 10 patients with low risk, and rarely classified a patient with moderate risk. RetinaVLM-Specialist was the best peforming VLM, and was able to detect 23 out of the 29 high-risk cases that require immediate treatment. At the same time, RetinaVLM's false discovery rate, defined as the ratio of the number of false positives over the number of predicted positives, of 42.5% is substantially lower than that of opticians at 69.5%. Owing to their ability to better differentiate moderate from high-risk cases, the human ophthalmologists had the lowest false discovery rate of 9.1%, although they simultaneously missed three more cases in urgent need for treatment. A full table of F1 scores for this task, including a comparison to a standard image-only classification model, are shown in Supplementary Fig. 6a.

In practice, referral letters should communicate the reason for referral by citing suspected abnormalities in the OCT image that can inform the ophthalmologist's initial diagnostic plan. As in the conciseness study in Fig. 4, RetinaVLM-Specialist sometimes documents the presence of small biomarkers that cannot be found in the image. More often, RetinaVLM-Specialist wrote an accurate imaging report but did not accurately follow the complex set of referral guidelines provided in the instruction. This led RetinaVLM-Specialist to incorrectly recommend that 17 of the moderate-risk patients potentially require treatment. However, this occurred less for the 25 low-risk patients, where RetinaVLM-Specialist correctly identified patients with little or no abnormalities, which are often referred to the treatment clinic for a second opinion by non-specialists (samples 1 and 3 in Fig. 5d). Crucially, we find that RetinaVLM-Specialist is effective in the detection of intraretinal cysts and fluid that differentiate high-risk from moderate-risk patients (samples 4 and 5). Four representative examples of success and failure cases of RetinaVLM-Specialist are shown in Supplementary Fig. 5b.

RetinaVLM accurately detects imaging biomarkers to make recommendations

It is important that clinical decision makers can provide evidence for their recommendations. Disease staging reports and written referral recommendations commonly contain descriptions of the most salient biomarkers that were detected in the scan. We tested the ability of four VLMs to correctly identify the presence or absence of 10 different biomarkers related to AMD. To this end, all VLMs were tasked with writing reports for 396 OCT images that conclude by stating the presence or absence of the biomarker in question (see Fig. 6a). The VLMs predictions were compared against the ground truth labels obtained from junior ophthalmologists. The instruction used to generate these biomarker focused reports is provided in the "Report generation for biomarker analysis" section.

We find that RetinaVLM-Specialist outperforms both LLaVA-Med and Med-Flamingo in the detection of seven out of the ten of main biomarkers related to AMD (see Fig. 6b). In overall performance, RetinaVLM-Specialist performed similarly well as ChatGPT-4o. In general, biomarkers that were more severe, larger, and more numerous were detected with higher accuracy by RetinaVLM-Specialist than less advanced presentations (see Fig. 6c). Most of the smaller biomarkers, such as small amounts of intraretinal fluid, drusen and hyperreflective foci, which can be as small as 30 µm in size[29], were detected with lower sensitivity. Overall, clinically important hallmarks of late AMD were detected with a very high sensitivity. Large volumes of subretinal and intraretinal fluid were detected in 80% and 78% of cases, respectively, and severe levels of hypertransmission in 84% of cases. We believe that stronger image encoders capable of detecting smaller features are necessary to improve the performance of RetinaVLM-Specialist on the detection of smaller biomarkers.

Finally, we compute saliency maps which highlight regions of the image that were most important to the model in writing specific passages of the report (see Fig. 6d). Qualitatively, we find that RetinaVLM-Specialist attends to relevant regions of the image containing fluid, hypertransmission and retinal pigment epithelium (RPE) irregularities in order to compose passages related to biomarkers and disease stage. Rather than performing biomarker segmentation, for which standard deep segmentation models by Isensee et al.[30] would be a more appropriate tool, these maps provide some explanation as to the choice of words and report passages made by the model. We calculate these using Grad-CAM[31] as described in the "Computing language-based image saliency maps" section, and provide four additional examples in Supplementary Fig. 8.

Discussion

In this study we have presented a curriculum-based approach for the specialization of medical vision-language models that directly leverages the knowledge of domain experts. Given a retinal OCT image, the resulting RetinaVLM-Specialist model generates accurate, detailed textual responses related to disease staging, referral or biomarker identification of AMD. While large foundation deep learning models have been employed for retinal image analysis before[22,32], our generative VLM is the first model that can flexibly process varied textual queries related to complex ophthalmological decisions and return detailed written responses. Through the use of language as primary communication medium, artificial intelligence systems are able to dynamically perform new tasks and meet the evolving requirements of image-based clinical decision makers.

In extensive experiments, RetinaVLM-Specialist significantly outperformed ChatGPT-4o and two generative VLMs, Med-Flamingo and LLaVA-Med, designed for medical use. RetinaVLM-Base and ChatGPT-4o were more capable in classifying conversion to late stages of AMD than existing medical foundation models, Flamingo and LLaVA-Med. However, they also lacked the ability to identify subtle yet important differences between disease stages and patient risk. Specifically, we have shown that in disease staging, RetinaVLM-Specialist outperforms all baselines and is approaching the accuracy of junior ophthalmologists. Similarly, when testing the ability of VLMs to screen for high-risk patients, ChatGPT-4o outperforms non-specialist opticians on aggregate but significantly underperforms compared to RetinaVLM-Specialist and junior ophthalmologists. In comparison, RetinaVLM-Specialist's reports reduced the number of incorrect urgent referrals by almost four times compared to opticians and had higher recall for urgent referrals than junior ophthalmologists. Finally, RetinaVLM is able to reinforce its decisions by citing observable biomarkers within the written report, and highlighting their corresponding regions within the image.

We postulate that the poor performance of ChatGPT-4o and medical foundation models alike stems from their lack of detailed knowledge related retinal OCT and AMD. Current VLMs including ChatGPT-4o and LLaVA-Med are trained on broad, unstructured datasets that are extracted from medical textbooks, scientific publications or social media posts of healthcare professional[9]. In the United Kingdom and the United States, clinical trainees aspiring to become specialists must undergo up to ten years of post-graduate training to obtain the grade of a board-certified consultant. Classifying AMD requires identifying subtle yet crucial differences between disease stages, particularly in biomarkers such as hypertransmission, drusenoid versus fibrous pigment epithelial detachments, and subretinal hyperreflective material — none of which ChatGPT-4o reported on. While ChatGPT-4o is capable of explaining these differences in language, our analysis indicates it cannot yet make these distinctions in images. This reinforces our insight that the paired image-text training data of current foundation VLMs lacks specialist and experiential knowledge, hindering their effective application to real-world clinical tasks.

A core innovation of our work was the creation of a dedicated training curriculum that specializes VLMs in image-based clinical decision making. Analogously to current medical education, this curriculum deconstructs clinical problems into sets of mandatory capabilities required for their resolution and selectively trains VLMs in these skills. To this end, we

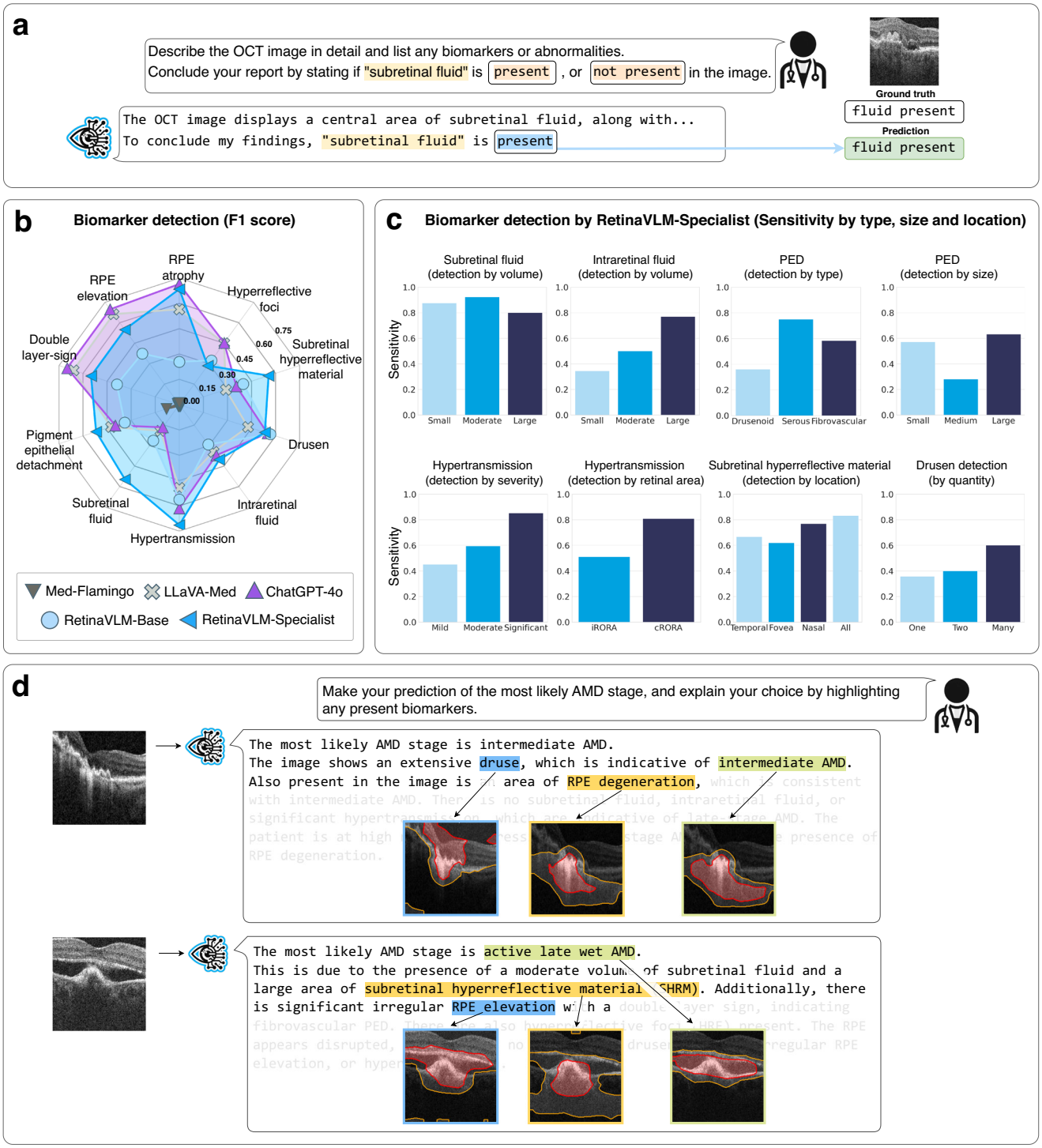


Fig. 6 | Comparison of vision-language models in analyzing imaging biomarkers. **a** We test four VLMs on their ability to describe the presence of 10 important imaging biomarkers of AMD. **b** Overall detection accuracy was computed using F1 scores. **c** Detection sensitivity for each level of biomarker severity for the most important biomarkers. **d** Regions highlighted by RetinaVLM-Specialist that correspond with different biomarkers and disease stage assessments in its written report. Regions with greater than 25% and 50% importance are highlighted by yellow and red contours, respectively. Four additional examples are shown in Supplementary Fig. 8.

obtained a large number of tabular reports by processing of retrospectively collected clinical data using advanced algorithms. Additionally, we tasked ophthalmologists to produce a limited number of highly specific textual reports. In total, our curriculum comprises 41,926 OCT images with 479,710 corresponding visual questions and answers. While still modest in size compared to substantially larger foundation datasets, we believe such curated needs-driven approaches are required to deploy language models specialist healthcare. In a similar vein, leading technology companies in artificial intelligence have also started to look beyond the internet's image and text data to source specialized training data to train LLMs and VLMs in disciplines such as computer programming, journalism, mathematics[33,34]. Future work may study specializing generalist models such as ChatGPT-4o by finetuning them on specific, high-quality curricula such as the one introduced in this study.

In the following, we discuss the technical limitations of our work. Naturally, the quality of our curriculum depends on the underlying imaging reports, and in particular those used to create curriculum part 2. The majority of the reports used to create RetinaVLM-Specialist were written by a junior ophthalmologist with three years experience. The remaining reports were written by an ophthalmologist with ten years of experience, and their reports were more comprehensive. While reports written by both were found to be of high quality, future work could evaluate the benefits of a third training curriculum derived by senior ophthalmologists.

By updating instruction sets, VLMs have the potential to be more adaptable than image-only deep learning approaches to differences in clinical practice between countries. However, it is important to note that our curriculum was derived using UK-specific clinical definitions and workflows. As a result, the errors made by RetinaVLM were found to reflect differences in the biomarker and staging definitions used by the UK-based and Austria-based ophthalmologists involved in the testing of the model. In particular, the model was found to be conservative in classifying fibrovascular features that upgrade intermediate AMD to inactive late wet AMD. Thus, to match or surpass the performance of the average junior ophthalmologist, we also aim to establish a consistent cohort of ophthalmologists for training and testing the model.

Similarly, we observed a discrepancy in image interpretation between junior and senior ophthalmologists. Junior ophthalmologists did not recommend patients for referral if it was likely that the retinal fluid observed was caused by traction rather than neovascularization, as it is not treatable with antiangiogenic drugs. Conversely, the senior ophthalmologist preferred that these patients be still referred for immediate assessment to rule out neovascularization. RetinaVLM was explicitly instructed to refer patients with any sign of fluid of any cause, and correctly referred more patients as a result.

Another technical limitation of our approach was the sensitivity of the LLM generating the question-answer pairs to the specifics of its instructions. Extensive trial and error were required to arrive at several instructions, listed in the Supplementary material, that resulted in diverse sets of high quality question-answer pairs. We discern that all aspects of dataset creation - deciding on the required capabilities, collecting specialized annotations and converting these to question-answer pairs - should be formalized to systematically compare different approaches and ultimately scale dataset curation in the future.

RetinaVLM also inherits some of the fundamental limitations of language models. LLMs are prone to confidently present false or fabricated information, termed hallucinations, which has been identified as problematic in medical contexts[13,14]. This phenomenon has also been observed in VLMs, where the generated text does not relate to any object or feature that is observable in the image[35,36]. Similarly, in several cases RetinaVLM was observed to hallucinate the presence of retinal fluid and consequently report a more advanced AMD stages than was necessary. RetinaVLM's output was also sensitive to the wording of questions and instructions. While this had little impact on our qualitative analysis, extensive trial and error was necessary to ensure that RetinaVLM responded with one of the provided options in the quantitative analyses.

We now outline factors limiting the clinical applicability of the current iteration of RetinaVLM, and propose directions for addressing these in future. Currently, RetinaVLM processes a single two-dimensional OCT image from one type of OCT scanner. In ophthalmological practice, decisions are made based on three-dimensional images from multiple time points, although many recent studies on the use of foundation models in ophthalmology also analyze two-dimensional images[22]. We mitigated the impact of this discrepancy on our study by tasking ophthalmologists to select the most relevant two-dimensional slice of the imaged volume before proceeding with the referral decision. In the future, a more sophisticated vision encoder, which is able to handle three-dimensional data from diverse OCT imaging devices, could be integrated with RetinaVLM. Our results also indicate that stronger image encoders capable of detecting smaller features are necessary to improve the performance of RetinaVLM-Specialist on the detection of smaller biomarkers. Similarly, one may opt to incorporate multimodal information, such as health questionnaires, clinical tests or the patient's medical history, into the decision making process[37]. The fundamental model architecture and training would remain similar, but the level of reasoning required for differential diagnosis across multiple scans would potentially increase.

Similarly, we exclusively trained RetinaVLM for the management of a single retinal disease, AMD, ignoring other retinal pathologies such as diabetic retinopathy or glaucoma, or imaging modalities, such as color fundus photography[38,39]. While this enabled us to explore the potential to encode advanced clinical levels of specialism into VLMs at depth, it would severely limit the applicability of the current version of our models for general ocular screening. It also led the model to classify all cases of retinal fluid as wet AMD, where a few cases showed evidence of diabetic retinopathy and other retinal conditions.

In order to ready VLMs for deployment for ocular screening, future work would need to extend RetinaVLM's curriculum with domain experts from an expanded range of ophthalmological conditions. This requires costly and time-consuming curation of specialized training datasets by medical experts. However, we believe that this is a necessary investment as routine clinical skills and patient management protocols are rarely documented in existing datasets used to train AI models.

Overall, our results indicate that merely increasing the scale of training datasets is insufficient for the development of VLMs with real-world clinical utility. Instead, medical VLMs require high-quality data directly related to the challenges faced by clinicians in their daily practice. We believe our proposed curriculum-based approach provides a blueprint for specializing VLMs that generate true value in healthcare.

Methods

Retinal image dataset curation

We use two retinal OCT datasets in this study. The first, described in the "Retrospective patient dataset" section, contains a cohort of patients with AMD collected retrospectively at the Southampton Eye Unit. The second, described in the "External testing dataset of referred patients" section, contains scans of the initial visits of patients referred, primarily by opticians, to the Southampton Eye Unit. The curation and use of this data is summarized in a flowchart diagram in Supplementary Fig. 11.

All data was collected in the scope of the PINNACLE study (ClinicalTrials.gov NCT04269304), which received approval from the East Midlands-Leicester Central Research Ethics Committee in the United Kingdom (ref. 19/EM/0163) and the institutional review boards of all participating institutions. It complies with the principles of Good Clinical Practice and the Declaration of Helsinki. Informed consent procedures were followed according to the principles of Good Clinical Practice and the Declaration of Helsinki. All images were captured using Topcon 3D OCT scanners (Topcon Corporation, Tokyo, Japan). Both datasets contain images of size 416×512 with a pixel size of $3.5 \times 11.7\ \mu m^2$.

Retrospective patient dataset

The retrospective dataset contains 44,633 OCT images from 6152 eyes belonging to 3468 patients, collected over eight years, between 2012 and 2020, at the Southampton Eye Unit and curated by the PINNACLE consortium. For each OCT scan we use the mediolateral 2D slice centered at the fovea. We then designated 41,926 of the 44,633 OCT images from 5547 eyes of 3057 patients for training purposes. Additionally, we reserved 2,311 images from 326 eyes of 187 patients for validation, and 396 images from 279 eyes of 224 patients for testing. We ensured that images from each patient do not appear in more than one of the training, validation or test sets. The training set was used to create both curriculum parts 1 and 2, detailed in Sections Report curation and question-answer pair generation (curriculum part 1) and Report curation and question-answer pair generation (curriculum part 2). The retrospective test set was used to evaluate the resulting model in the disease staging and biomarker verification section in the Results.

To evaluate the accuracy of image report conclusions, we used a multi-rater process for curating high-quality labels. We now expand on this by

detailing the breakdown of disease stage definitions used in this study. AMD remains a relatively poorly understood disease, with multiple grading systems that vary in stage classification and definitions[40–44]. Its classification remains an active area of research[44]. Classifying a retinal OCT image into a single disease stage is challenging, as overlapping features can indicate multiple AMD stages, often requiring ophthalmologists to make intuitive assessments beyond strict criteria. In this study, the ophthalmologists reported classifying healthy and early AMD based on the presence or absence of small drusen, and intermediate AMD by medium, or intermediate, to large drusen[41]. Late dry AMD was identified by atrophy of the retinal pigment epithelium, evidenced by hypertransmission. Presence of any subretinal hyperreflective material or other scarring advanced the classification to inactive wet AMD. Finally, presence of any subretinal or intraretinal fluid upgraded the image-based classification to active late wet AMD. These classifications yielded a testing set of 276 images that labeled 36 as healthy, 18 as early, 64 as intermediate, 25 as late dry, 75 as inactive late wet and 58 as active late wet AMD.

Each image was labeled with the presence or absence of 10 biomarkers, and an additional 21 labels that record their size, number, and other applicable attributes. Due to the substantially increased number of variables compared to disease staging, each image was labeled only once by one of the six junior ophthalmologists. This process yielded a testing set of 396 images that labeled 107 as evidencing drusen, 70 with pigment epithelial detachments, 150 with pigment epithelial elevation, 27 with double-layer sign, 122 with hypertransmission, 155 with atrophic/degenerated pigment epithelium layers, 60 with subretinal hyperreflective material, 81 with hyperreflective foci, 25 with subretinal fluid, and 62 with intraretinal fluid.

External testing dataset of referred patients

We also collected an external dataset of 95 patients that were referred primarily by opticians for urgent care at the Southampton Eye Unit between 02/2023 and 12/2023. None had yet received treatment for AMD, and mostly had no AMD, intermediate AMD or small features related to active wet AMD. This represents a distribution shift from the retrospective cohort, where many patients had already received treatment for AMD and were in the inactive late wet stage of AMD. As such, it enabled us to estimate the robustness of both variants of RetinaVLM to shifts in patient population. This dataset was not used for model training and was reserved for testing VLMs on patient referral as shown in Fig. 5.

For each patient we sourced scans of both their left and right eye that were acquired on their first visit to the clinic. We also collected the originally issued letter of referral, as depicted in Fig. 5d. Then, two junior ophthalmologists independently reviewed the 3D OCT volumes of both eyes to label the patient's risk and decide from three levels of increasing referral referral urgency. To help standardize their labels, the ophthalmologists referred to a set of agreed patient referral guidelines. For full documentation see `<PatientReferralGuidelines>` in Supplementary Fig. 10c.

After labeling the level of referral urgency, the ophthalmologists then selected the image slice that most supported their assessment of the patient's risk. In healthy patients where both volumes contained no pathological signs in any of the image slices, they were instructed to select the mediolateral fovea-centered 2D slice from one of the two volumes. Finally, any inter-rater disagreements were then arbitrated by a panel involving the two senior ophthalmologists. Of the 95 images in the external cohort, 25 did not evidence need of referral, 41 indicated the patient was at moderate risk and needed general attention by a specialist, while only 29 indicated the patient needed urgent referral.

RetinaVLM vision-language model architecture

RetinaVLM follows the architectural design of MiniGPT4[24]. It consists of two main components: an ophthalmological vision encoder and a generative LLM (see Supplementary Fig. 9). For the ophthalmological vision encoder, we adopt a Resnet50 convolutional neural network with over 23 million parameters which was pre-trained with self-supervised contrastive learning on the 41,926 OCT images from the train set of the retrospective cohort. Specifically, it was trained with Bootstrap Your Own Latent (BYOL)[45] using the same implementation details as the standard contrastive approaches in[21], which consistently performed on par with RETFound[22] on data from the Southampton Eye Unit. This vision encoder projects each 192×192 input image to a set of spatially arranged 6×6 visual embeddings, which are extracted from the last layer before global average pooling. Each embedding has a dimension of $h_{img} = 2048$. They also have a receptive field of size 336, so each embedding contains global knowledge of the image that is contextualized at its local position.

For the LLM, we employ the 8 billion parameter instruction-tuned Llama3 model by Meta[23,46] as the generative LLM, which was was the most performant openly available model at the time of our study. LLama3 uses an embedding dimension of $h_{lang} = 4096$.

The ophthalmological vision encoder provides visual information regarding the OCT image to the LLM via an adapter. The adapter is a linear layer of size $h_{img} \times h_{lang}$ that processes visual information for use by the LLM. Specifically, it does so by independently mapping each of the visual embeddings, applying a linear transformation via matrix multiplication, into language embeddings used by the LLM. The design and application of the adapter follows the design used in MiniGPT4[24], and results in an adapter with over 8 million parameters.

Foundation vision-language models

We used the two most widely adopted foundation vision-language models for medical applications at the time of this study[9,10]. They were both trained on large biomedical datasets sourced from the Internet, and have been applied in chest x-ray[47]. The first, Med-Flamingo[10], which was built on Flamingo[48] and finetuned on image and text data from medical textbooks and the PubMed Central Open Access (PMC-OA) dataset[49]. The second, LLaVA-Med[9], developed by Microsoft, is a VLM built on LLaVA[18] and finetuned to follow textual instructions regarding a broad range of biomedical images contained in PubMed Central 15M (PMC-15M)[50]. The training sets of both models contain retinal OCT images and associated text. As they were trained as generalist models on various imaging modalities, they were both purportedly capable of interpreting retinal OCT images. When provided a retinal OCT image we found both could correctly identify its modality and begin generating diagnoses, without being provided any prior information. More recently, other generative vision-language models for ophthalmology have been introduced but were either not designed for OCT imaging nor finetuned with instruction-tuning[51,52]. Our third baseline constitutes OpenAI's GPT-4o model (endpoint 'gpt-4o-2024-05-13'). Unlike the two aforementioned medical VLMs, GPT-4o is a generalist model.

For Med-Flamingo, we then provide instructions using the following template provided in their code, replacing `{question}` with the instruction text:

```
You are a helpful medical assistant. You are being provided with images, a question about the image and an answer. Follow the examples and answer the last question. <image>Question: {question} Answer:
```

Similarly, for LLaVA-Med we use the following conversation template provided by the developers:

```
A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions.###Human: {question}###Assistant:
```

Similarly, for the ChatGPT-4o API we provide the following system prompt:

```
You are an intelligent and helpful assistant. You follow all the instructions in the user prompt, and answer all the questions they ask for.
```

Report curation and question-answer pair generation (curriculum part 1)

To create the tabular reports for the first part of the curriculum we used a cluster-based approach to efficiently label the 41,926 training images with biomarker annotations[53]. This training cohort includes 25,825 female and 16,101 male patients with an average age of 81.8 years, with lower (Q1) and upper (Q3) quartiles of of 77.0 and 88.0 years.

Contrastive learning is used to extract self-supervised features from the dataset. The dataset is then partitioned into 40 clusters of images that share common features. Labels are then assigned to these clusters by senior ophthalmologists. To this end, 20 images from each cluster were reviewed by senior ophthalmologists. If the majority of the images exhibited common features, such as 'large drusen' or 'subretinal fluid', these labels were assigned to the entire cluster. A review quantifying the accuracy of each cluster description is performed in Supplementary Section A.4.

These labels were used in in combination with the patient's age, sex and their functional visual acuity score (measured on a LogMAR chart and converted to Letter score) to create the tabular reports. We also included the quality index of the image using an intensity-histogram-based approach[54]. We found that labeling the dataset's quality index quartiles as 'very poor', 'ok,' 'good,' and 'excellent' effectively captured the characteristics of each quartile. Additionally, the reports list three biomarkers that are stated as not being present. These are drawn from a distribution of all biomarkers, weighted by their prevalence in the dataset, that were not among the cluster labels for that image. Counts of the prevalence of each tabular variable among the training images are shown in Supplementary Fig. 1a, b, and a sample of four tabular reports they result in are shown in Supplementary Fig. 1c.

To generate question-answer pairs from the large volume of tabular reports we used an LLM available for download and local use. We chose WizardLLM-70B, though any freely-available LLM with instruction-following capabilities may be used for this purpose. This model was chosen as an open-weights model with strong performance in instruction following and general knowledge at the time of dataset creation. We then used the instruction template detailed in the "Question-answer generation prompt (curriculum part 1)" section in the Supplementary material. This resulted in a total of 408,505 question-answer pairs, or an average of 9.75 question-answer pairs per report depending on its complexity. Examples of the question-answers pairs generated by this approach are shown in the 'curriculum part 1' section of Supplementary Fig. 4a, b.

Report curation and question-answer pair generation (curriculum part 2)

The second part of the curriculum involves further turning on a subset of 330 images from the training curriculum in part 1. This training cohort includes 205 female and 125 male patients with an average age of 80.8 years, with lower (Q1) and upper (Q3) quartiles of of 75.0 and 87.0 years.

Unlike curriculum part 1, images used in curriculum part 2 were annotated with high quality reports manually curated by retinal specialists. Specifically, two junior ophthalmologists were tasked with describing the main pathological biomarkers and diagnoses related to AMD, while also noting any other observation regarding the retinal anatomy. This yielded high-quality textual reports that go beyond the short notes that ophthalmologists typically write in clinical routine. The first junior ophthalmologist, with three years of experience specializing in ophthalmology, wrote the majority of 244 reports (see Supplementary Fig. 2a). While these were highly accurate, they were less comprehensive in their analysis than the remaining 86 reports written by the junior ophthalmologist with 10 years of experience (see Supplementary Fig. 2b).

Simultaneously, the same two junior ophthalmologists used the same methodology to produce 28 reports on images from the test set. In our analysis, these were found to be of high quality by the two senior ophthalmologists (see Fig. 4). As they were representative of the 330 reports collected on the training set, the senior ophthalmologists concluded that these results provide sufficient quality assurance for the reports used to generate curriculum part 2.

After the imaging reports were curated, we used an external LLM with 10 different instructions to generate up to 230 questions per image. The exact instructions used are documented in the "Question-answer generation prompts (curriculum part 2)" section in the Supplementary material. We take two preliminary steps before providing the instruction to the LLM. We firstly replace the `<ReportText>` identifier with the raw text of the image report. Additionally, many of the QA generation instructions make references to the guidelines that describe the mandatory capabilities of image-based clinical decision makers with regard to disease staging and patient referral for patients with AMD. The second step involves replacing any reference to the `<ObservationGuidelines>`, `<DiseaseStagingGuidelines>` or `<PatientReferralGuidelines>` by the text of the corresponding guidelines (documented in Supplementary Fig. 10). These guidelines were verified by senior ophthalmologists, and were instrumental for the generation of questions with improved diversity and coverage, and also for creating questions about biomarkers that were absent in the image (and typically not mentioned in the report).

The smaller number of reports in the advanced curriculum permitted the use of the more performant proprietary models for generating question-answer pairs. We used the 'gpt-4o' API endpoint from OpenAI. An example interaction with 'gpt-4o' for generating these question-answer pairs is shown in Supplementary Fig. 3. A sample of question-answers yielded by this approach are shown in the 'part 2' section of both Supplementary Fig. 4a, b.

RetinaVLM model inference

RetinaVLM processes retinal OCT images and textual instructions. To begin building the input provided to the model, we first set the system prompt of the constituent LLM to:

```
You are a helpful ophthalmological specialist
chatbot capable of interpreting retinal OCT
images.
```

We then begin the instruction that will be provided to the LLM with the following line:

`Here is an encoding of a retinal OCT image <Img><ImageHere></Img>\n`

For each image in the batch, we next add the text of the specific question to the instruction. For example:

```
Describe the OCT image in detail. Does it show evi
dence of retinal fluid?
```

To complete the textual input to the model, we populate the LLM's conversation template with this instruction and include the question's answer exclusively during training. This forms the full textual input, which we project to the embedding space of the LLM using its pre-existing tokenizer and pretrained embedding layer.

After embedding the input text, we process the retinal OCT image in question. To this end we downsample the image by a factor of 2 from 416×512 to 208×256 pixels. We then crop the image centrally in testing, and augment each image using the finetuning protocol outlined in ref. 21 in training. This results in a processed OCT image of size 192×192.

Next, we use the ophthalmic vision encoder E to extract the 6×6 vision feature embeddings from the image. After flattening these, we apply the adapter to each embedding separately. This projects the visual embeddings to the input embedding space of the LLM. To create the final set of embeddings that are provided to the LLM, we replace the textual embeddings corresponding to the `<ImageHere>` phrase with the 36 adapted vision embeddings from the actual retinal OCT image.

Finally, the resulting sequence of language embeddings and adapted visual embeddings are passed together through the frozen LLM. This yields a

list of predicted token logits with the same length as the input sequence. We then compute the causal language modeling loss between the predicted answer logits and the ground truth answer tokens. We then optimize the adapter to minimize this loss.

Crucially, when training RetinaVLM we keep both the vision encoder and LLM *frozen*. That is, they are not updated during the entire training process. This is key to preserving RetinaVLM's inheritance of the pretrained LLM's language and reasoning capabilities. These enable RetinaVLM to handle versatile questions and instructions beyond those encountered in the curriculum during training. Thus, during training, we only update the adapter that feeds visual information regarding the OCT image to the LLM.

Training RetinaVLM on curriculum parts 1 and 2

RetinaVLM is trained sequentially on curriculum part 1 and part 2. After randomly initializing the adapter we first train the model on the 408,545 question-answer pairs regarding the 41,926 images in curriculum part 1 (introduction to retina) to obtain *RetinaVLM-Base*. The then continue to finetune the model on the 71,165 questions and answers regarding the 330 images in curriculum part 2 (advanced retinal specialism), resulting in the final *RetinaVLM-Specialist* model. For each of the curriculum parts we train RetinaVLM for 100,000 steps using a batch size of 12. In each step we randomly select one question-answer pair per image from the current curriculum. To update the network we use the AdamW optimizer with a learning rate of 10^{-4}, and $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

Training on the two curriculum parts separately is important for assessing the accumulative benefits of progressively specialist training in two ways. Firstly, this enables us to measure the benefits of training on free-text medical reports over tabular data. Secondly, during development we found that reserving the highest quality data for the final training stage, curriculum part 2, improved the performance of the resulting model. This strategy is standard practice in the development of LLM-based models[17].

Experimental setup

After VLMs have been trained on image and text datasets, they can be used to generate responses to new questions and images. To use the baseline foundation VLMs for testing we take the central crop in each 416 × 512 OCT image, resulting in an image of size 384 × 384. After repeating this image along the color dimension, ChatGPT-4o accepts the resulting 3 × 384 × 384 RGB image, while the Med-Flamingo and LLaVA-Med baselines require we downsample this to 3 × 224 × 224. To provide the image to the RetinaVLM variants, we first downsample the 416 × 512 image by a factor of two to 208 × 256, before taking a central crop resulting in an image of size 192 × 192.

We then employ the same method with all VLMs for generating responses to instructions. Provided the image and textual instruction, we build the output sequence of tokens by repeatedly appending the token to the output assigned the highest probability by the VLM. This is equivalent to using a temperature parameter set to 0, and is the standard approach for generating the most accurate and least creative output from LLM-based models. All other generation parameters, such as repetition penalty, were not used. This process is repeated until a stop token is generated, signaling the end of the VLM's response. The model's tokenizer is then used to convert the numeric output tokens to the final free text output.

Our entire evaluation was conducted in *zero-shot*, that is, after training on curriculum part 1 and part 2 RetinaVLM requires no further finetuning in order to perform tasks related to disease staging, patient referral and biomarker analysis. Instead, for each test we designed a specific instruction that was provided to all VLMs to generate the application-specific reports that were used in our analyses. These instructions were derived through experimentation with all VLMs on the validation set, and are documented in the following four subsections. To inform their design, we assessed the effectiveness of each prompt by assessing the quality of outputs it produced in all VLMs on a subset of 30 images from the validation set. This led to two observations. Firstly, that all VLMs produced improved responses when prompted to first detail its 'chain-of-thought'[55]. This led us to request that the model first describe the OCT image in detail before making any recommendations. Secondly, that simpler, more direct prompts resulted in all VLMs making fewer errors in following our subsequent, test-specific instructions.

Report generation for disease staging

The following instruction was given to all VLMs to and obtain the reports of the 276 test images that were analyzed in Fig. 3. The instruction requests VLMs to begin by describing the image, and then deduce the most advanced disease stage:

```
Describe the OCT image in detail and list any bio
markers or abnormalities, including the most
likely AMD stage of the patient.

Then, based on those observations, state if the
patient's most advanced AMD stage is "healthy',
"early', "intermediate', "late dry', "late wet
(inactive)' or "late wet (active)'?
```

After the VLM generated its report (using a maximum of 500 tokens), we appended the following incomplete sentence to its output:

```
Based off the image and those findings, the patient's
most advanced AMD stage is
```

which the VLM them completes using up to 300 tokens. From these tokens, we extracted the final disease staging prediction by searching for the first instance of any of the listed disease stages. This post-processing step is only necessary for quantitative tests of accuracy, as it enables the reliable extraction of the disease stage from the free text report. In cases where the VLM discusses multiple disease stages, such as in '`more advanced than early AMD, and is intermediate AMD as there is no evidence of late wet AMD`', the disease stage was manually extracted. In cases where no disease stage was provided or could be extracted manually, this counted as an 'Invalid response'.

Report generation for qualitative evaluation

For the qualitative evaluation by the senior ophthalmologists, we randomly selected 28 of the test images from the retrospective dataset, and tasked the two junior ophthalmologists with annotating the images. We provided the following instruction to RetinaVLM-Specialist and LLaVA-Med to generate their image reports:

```
Write an extensive report describing the OCT
image, noting any biomarkers or abnormalities
related to AMD, and their qualities. Also comment
on which biomarkers are absent.

Finally, based on the image and these findings,
your report should estimate the AMD disease stage
of the patient.

You should not include any patient referral
recommendations in your report, but you can com
ment if they need treatment with anti-vegf.
```

We found that ChatGPT-4o tended to write excessively long reports. To improve the performance of ChatGPT-4o in direct evaluations by senior ophthalmologists, we requested a '`brief`' rather than an '`extensive`' report, and appended the following text to the original instruction: '`Your`

report should be no longer than 120 words long'. This helped make reports more concise with no observable loss in accuracy, though they still exceeded the length of reports written by junior ophthalmologists and RetinaVLM-Specialist.

We then assigned 14 of the images and the corresponding 42 reports by ChatGPT-4o, RetinaVLM-Specialist and the junior ophthalmologists, to each of the two senior ophthalmologists, who evaluated them independently according to the correctness, completeness and conciseness. These criteria, also documented in Fig. 4, were:

- Correctness - The report is accurate in its main observations and conclusions regarding the image
- Completeness - The report contains all relevant observations and conclusions that can be inferred from the image
- Conciseness - The report does not make observations and conclusions that are not supported by, or not seen in, the image

Report generation for patient referral

The following instruction was given to all VLMs to generate reports that focus on patient referral recommendations, which are analyzed in Fig. 3. This instruction was run for the 95 referral images, introduced in the "External testing dataset of referred patients" section. In order to accurately convey the specific requirements of the wet AMD treatment clinic, we provided the comprehensive referral protocol used by the senior ophthalmologists in the instruction:

Do not provide a disease stage, or referral recommendation yet.
Being seen by a specialist at the Southampton clinic:

A. The Southampton clinic requires that patients with any sign of intraretinal fluid, any sign of subretinal fluid, or any sign of cyst(s), MUST be seen by a specialist at the Southampton clinic within the next two weeks.
B. The Southampton clinic requires that patients who do not have any sign of intraretinal fluid, any sign of subretinal fluid, or any sign of cyst(s), but do have some biomarkers of early or intermediate AMD, should be seen by a specialist at the Southampton clinic for routine referral.
C. The Southampton clinic requires that patients who do not have any sign of intraretinal fluid, any sign of subretinal fluid, or any sign of cyst(s), but do have medium to large drusen, drusenoid PED, hypertransmission or atrophy, should be seen by a specialist at the Southampton clinic for routine referral.
D. The Southampton clinic does not need to see patients who have no biomarkers and healthy retinas at all.

```
Southampton specialist visit: Next, tell me if
your initial report of the OCT image indicates
that the patient should be seen by a specialist at
the Southampton clinic "within the next two
weeks", to be seen "within 18 weeks (routine
referral)", or "not be seen" at all?
```

As before, after the VLM generated its report (using a maximum of 500 tokens), we added the following incomplete sentence to its output:

```
My report indicates that the patient
```

which is then completed by the VLM for up to a maximum of 300 tokens. We then searched these tokens for the first instance of one of the three levels of referral urgency, 'within the next two weeks', 'within 18 weeks (routine referral)' or 'not be seen', in the VLM's output report. In cases where different wording is used, but with identical meaning, the VLM's prediction is extracted manually. In the remainder of cases, it is listed as an 'Invalid response'.

Report generation for biomarker analysis

The following instruction was given to all VLMs to generate reports that conclude the presence of absence of the 10 different biomarkers, evaluated on the 396 test images in Fig. 6:

```
Describe the OCT image in detail and list all bio
markers or abnormalities. Detail if there are any
signs indicating that biomarker might be present,
even if there is only a small amount.

Finally, conclude your findings by telling me if
biomarker article "not present", or if poten
tially any amount of biomarker article "present"
in the OCT image.
```

For each of the 10 biomarkers, the phrase {biomarker} was replaced by the actual biomarker name (such as 'subretinal fluid'), and the {article} replaced by 'is' for singular biomarkers or 'are' for plural biomarkers (such as drusen). After the VLM generated its report (using a maximum of 500 tokens), we appended the following incomplete sentence to its output:

```
To conclude these findings, in the OCT image {bio
marker} {article}
```

which the VLM then completes using up to 300 tokens. We then searched for the first instance of not present or present to extract the model's prediction of the absence or presence of the biomarker, respectively. In cases where different wording is used, but with identical meaning, such as stating the biomarker was 'detected' rather than 'present', the VLM's prediction is extracted manually. In the remainder of cases, it is listed as an 'Invalid response'.

Computing language-based image saliency maps

We provide methodogolical details for the computation of the language-based saliency maps shown in Fig. 6d and Supplementary Fig. 8. With saliency maps we aim to identify which regions of the image were most relevant to certain passages, such as large subretinal fluid, of RetinaVLM-Specialist's responses. The most direct way to generate these visualizations to use attention maps, but we found Llama3's pretrained attention maps did not result in any meaningful saliency maps. To address this, we used Grad-CAM[31], a technique for highlighting the most relevant image regions to the prediction of an image classifier. Specifically, we apply Grad-CAM to the sum of the tokens in the output passage, effectively reframing the LLM as an image classifier where the output corresponds to a specific passage of the report. This allows us to generate saliency maps relative to that passage. A code implementation can be found at the repository referenced in the "Code availability" section.

Comparison to an image-only classification-based deep learning baseline

VLMs are an emerging technology that hold great potential to automate language-based reporting and decisions regarding medical images. To provide a comparison between VLMs and more established approaches to classification we use RETFound, which was pretrained on over 700,000 2D OCT images and exhibits strong performance when finetuned on as few as 100 images[21,22]. RETFound makes predictions from the image alone and

cannot interpret nor respond to written language. As an image encoder, RETFound instead outputs a single classification per image.

We compare the VLMs against RETFound in both the disease staging and patient referral tasks. To this end, we formulated disease staging as a six-way classification task using the same stages as the experiment shown in Fig. 3. Similarly, we formulated the patient referral task as a three-way classification task using the same referral levels as the experiment shown in Fig. 4. In both cases, we manually extract training labels from the imaging reports of the same 330 training images used in curriculum part 2 to train RetinaVLM-Specialist (see the "Retrospective patient dataset" section).

After formulating these tasks, we used the standard linear evaluation approach to evaluate the performance of the RETFound imaging encoder. This involves applying global average pooling to RETFound's output tokens to yield a feature vector of size 1024. We then train a linear layer that maps this feature vector to the output classes of the given classification problem. This was done using an AdamW optimizer with learning rate of $3{\cdot}10^{-4}$ for 5000 training steps. In both cases, performance converged and little to no overfit was observed on the validation set. We then selected the step with the best performance on the validation set for evaluation. Finally, this classifier is tested on the same arbitrated test labels as all VLMs in Figs. 3 and 4.

The results of both experiments are shown in Supplementary Fig. 6a. For the disease staging task we find that RETFound (0.63 F1) significantly outperforms the best performing foundation medical VLMs (0.11 F1) and RetinaVLM-Base (0.30 F1). Notably, it performs as well as RetinaVLM-Specialist (0.63 F1). However, on the patient referral task RETFound (0.37 F1) performs on par with the best foundation VLM (0.39 F1) and significantly worse than RetinaVLM-Specialist (0.67 F1).

The equal performance of RetinaVLM-Specialist and the image-only RETFound baseline in disease staging implies that both have reached an upper bound on performance for this task on the retrospective dataset. However, we observed relatively poor performance from the image-only baseline on the patient referral task, which we attribute to domain shift. Specifically, the retrospective dataset lacks images with intact retinal pigment epithelium layers that feature small fluid pockets, which represent many of the urgent referral cases in the referral dataset. This image-only model's ability to generalize well to these cases. While accurate patient referral is achievable using an image-only encoder like RETFound, addressing this domain shift would necessitate the collection of a new training dataset that is more representative of the referral cohort. In contrast, RetinaVLM-Specialist is better able to mitigate this domain shift by utilizing task-specific textual instructions that provide details about the target domain – namely, the patient referral protocol outlined in the "Report generation for patient referral" section.

Measurements of performance and statistical analysis

To calculate the performance of each VLM and retinal specialist in all multiple-choice question answering, we used the micro F1 score. This aggregates the total number of false positives (FP), false negatives (FN), true negatives (TN) and true positives (TP) over all classes before computing the F1 score using Eq. (1)

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{1}$$

In cases where the VLM returned an 'Invalid response' by failing to pick one of the listed options this was counted as a false negative for the ground truth class. After calculating the F1 score, we determined the 95%confidence interval through bootstrapping $N = 1000$ times with replacement.

Tests of significance (aggregated in Supplementary Fig. 6) were calculated using a two-sided McNemar's test[56]. This test assesses the difference in the number of correctly versus incorrectly predicted samples, focusing on cases where the models agree or disagree on the labels. A significant p-value from the McNemar test allows us to reject the null hypothesis that both models have identical classification performance. We then used the following notation to indicate levels of statistical significance: *** for $p \leq 0.001$, ** for $p \leq 0.01$, and * for $p \leq 0.05$ and 'ns' (not significant) for $p > 0.05$.

Computing hardware and software

We use Python 3.12.2 to conduct all model question-answer generation, VLM training, and VLM evaluation. To generate the question-answer pairs for curriculum part 1 we used 3 40GB NVIDIA A40 GPUs. For both training RetinaVLM and for evaluating all VLMs we use a single 80GB NVIDIA A100 GPU and PyTorch[57] version 2.1.2. Training RetinaVLM on takes 1 day on curriculum part 1, and another day on curriculum part 2. Llama3 was downloaded via Huggingface with model ID 'meta-llama/Meta-Llama-3-8B-Instruct'. The baseline VLM Med-Flamingo's code and model weights were installed following the instructions at https://github.com/fastscience-ai/MedFlamingo, and LLaVA-Med's from https://github.com/microsoft/LLaVA-Med. Confusion matrices and results calculations were computed with scikit-learn version 1.4.1 and numpy version 1.26.4. Figures and tables were created in draw.io v24.4.0 using plots generated by matplotlib version 3.8.4 and seaborn version 0.13.1. Grad-CAM was computed using grad-cam version 1.5.0. McNemar's tests of significance were calculated using statsmodels version 0.14.1.

Data availability

The datasets used and analyzed during the current study are currently being curated and maintained by the Vienna Reading Center on behalf of the PINNACLE consortium. The data will be made available to once the PINNACLE study concludes in 2026. A minimal dataset is available from the corresponding author before 2026[58] upon reasonable request. Moreover, the code repository includes a minimal dataset that can be used to interpret, verify and extend the research in the article.

Code availability

The code and guidelines used to create the question-answer pairs, train, and evaluate the models can be found at https://github.com/RobbieHolland/SpecialistVLMs. The code may be used to develop the models can be repurposed for other medical specialties. Moreover, we make all model weights for RetinaVLM-Base and RetinaVLM-Specialist openly available at https://huggingface.co/RobbieHolland/RetinaVLM. These are only intended for research purposes related to retinal OCT images.

Received: 20 November 2024; Accepted: 15 July 2025;
Published online: 19 August 2025

References

1. Moy, A. J. et al. Measurement of clinical documentation burden among physicians and nurses using electronic health records: a scoping review. *J. Am. Med. Inform. Assoc.* **28**, 998–1008 (2021).
2. Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical AI. *Nat. Med.* **28**, 1773–1784 (2022).
3. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
4. Rajpurkar, P. & Lungren, M. P. The current and future state of ai interpretation of medical images. *N. Engl. J. Med.* **388**, 1981–1990 (2023).
5. Zhang, Y., Jiang, H., Miura, Y., Manning, C. D. & Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, 2–25 (PMLR, 2022).
6. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual–language foundation model for pathology image analysis using medical twitter. *Nat. Med.* **29**, 2307–2316 (2023).
7. Lu, M. Y. et al. A visual-language foundation model for computational pathology. *Nat. Med.* **30**, 863–874 (2024).
8. Christensen, M., Vukadinovic, M., Yuan, N. & Ouyang, D. Vision–language foundation model for echocardiogram interpretation. *Nat. Med.* **30**, 1481–1488 (2024).

9. Li, C. et al. Llava-med: training a large language-and-vision assistant for biomedicine in one day. *Adv. Neural Inf. Proces. Syst.* **36** 28541–28564 (2024).
10. Moor, M. et al. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health*, 353–367 (PMLR, 2023).
11. Tu, T. et al. Towards generalist biomedical AI. *NEJM AI* **1**, AIoa2300138 (2024).
12. Kung, T. H. et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digital Health* **2**, e0000198 (2023).
13. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
14. Hager, P. et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* https://www.nature.com/articles/s41591-024-03097-1 (2024).
15. Bengio, Y., Louradour, J., Collobert, R. & Weston, J. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48 (ICML, 2009).
16. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
17. Wei, J. et al. Finetuned language models are zero-shot learners. In *Proceedings of the International Conference on Learning Representations*, (ICLR, 2022).
18. Liu, H., Li, C., Wu, Q. & Lee, Y. J. Visual instruction tuning. In *Advances in neural information processing systems*, 36 (NeurIPS, 2024).
19. Mitchell, P., Liew, G., Gopinath, B. & Wong, T. Y. Age-related macular degeneration. *Lancet* **392**, 1147–1159 (2018).
20. Wong, W. L. et al. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *Lancet Glob. Health* **2**, e106–e116 (2014).
21. Holland, R. et al. Metadata-enhanced contrastive learning from retinal optical coherence tomography images. *Med. Image Anal.* **97**, 103296 (2024).
22. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
23. Meta AI. Introducing Meta Llama 3: The most capable openly available LLM to date https://ai.meta.com/blog/meta-llama-3/ (2024). Accessed: 2024-06-25.
24. Zhu, D. et al. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *Proceedings of the International Conference on Learning Representations*, (ICLR, 2024).
25. Anderson, P. et al. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086 (IEEE, 2018).
26. Resnikoff, S., Felch, W., Gauthier, T.-M. & Spivey, B. The number of ophthalmologists in practice and training worldwide: a growing gap despite more than 200 000 practitioners. *Br. J. Ophthalmol.* **96**, 783–787 (2012).
27. OpenAI. Gpt-4o: Openai's multimodal language model. https://openai.com/index/hello-gpt-4o/ (2024). Accessed: 2025-02-17.
28. Van Veen, D. et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.* **30**, 1134–1142 (2024).
29. Fragiotta, S. et al. Significance of hyperreflective foci as an optical coherence tomography biomarker in retinal diseases: characterization and clinical implications. *J. Ophthalmol.* **2021**, 6096017 (2021).
30. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
31. Selvaraju, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626 (IEEE, 2017).
32. De Fauw, J. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
33. Sambasivan, N. et al. Everyone wants to do the model work, not the data work: data cascades in high-stakes AI. *SIGCHI, ACM,* https://research.google/pubs/everyone-wants-to-do-the-model-work-not-the-data-work-data-cascades-in-high-stakes-ai/ (2021).
34. Heikkilä, M. OpenAI's hunger for data is coming back to bite it. MIT Technology Review, https://www.technologyreview.com/2023/04/19/1071789/openais-hunger-for-data-is-coming-back-to-bite-it/ (2023).
35. Li, Y. et al. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, (EMNLP, Singapore, 2023).
36. Liu, H. et al. A survey on hallucination in large vision-language models. Preprint at https://arxiv.org/abs/2402.00253 (2024).
37. Baltrušaitis, T., Ahuja, C. & Morency, L.-P. Multimodal machine learning: a survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 423–443 (2018).
38. Flaxman, S. R. et al. Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis. *Lancet Glob. Health* **5**, e1221–e1234 (2017).
39. Abràmoff, M. D., Garvin, M. K. & Sonka, M. Retinal imaging and image analysis. *IEEE Rev. Biomed. Eng.* **3**, 169–208 (2010).
40. Bird, A. C. et al. An international classification and grading system for age-related maculopathy and age-related macular degeneration. *Surv. Ophthalmol.* **39**, 367–374 (1995).
41. Klein, R. et al. Harmonizing the classification of age-related macular degeneration in the three-continent amd consortium. *Ophthalmic Epidemiol.* **21**, 14–23 (2014).
42. Ferris, F. L. et al. A simplified severity scale for age-related macular degeneration. *Arch. Ophthalmol.* **123**, 1570–1574 (2005).
43. Ferris III, F. L. et al. Clinical classification of age-related macular degeneration. *Ophthalmology* **120**, 844–851 (2013).
44. Sadda, S. R. et al. Consensus definition for atrophy associated with age-related macular degeneration on oct: classification of atrophy report 3. *Ophthalmology* **125**, 537–548 (2018).
45. Grill, J.-B. et al. Bootstrap your own latent-a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **33**, 21271–21284 (2020).
46. AI@Meta. Llama 3 model card, https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md (2024).
47. Chen, Z. et al. Chexagent: towards a foundation model for chest x-ray interpretation. Preprint at https://arxiv.org/abs/2401.12208 (2024).
48. Alayrac, J.-B. et al. Flamingo: a visual language model for few-shot learning. *Adv. neural Inf. Process. Syst.* **35**, 23716–23736 (2022).
49. Lin, W. et al. PMC-CLIP: contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 525–536 (Springer, 2023).
50. Zhang, S. et al. Large-scale domain-specific pretraining for biomedical vision-language processing. Preprint at https://arxiv.org/abs/2303.00915 (2023).
51. Deng, Z. et al. Ophglm: an ophthalmology large language-and-vision assistant. *Artif. Intell. Med.* **157**, 103001 (2024).
52. Zhang, K. et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nat. Med.* **30**, 3129–3141 (2024).
53. Holland, R. et al. Deep-learning-based clustering of OCT images for biomarker discovery in age-related macular degeneration (PINNACLE study report 4). *Ophthalmol. Sci.* **4**, 100543 (2024).
54. Stein, D. M. et al. A new quality assessment parameter for optical coherence tomography. *Br. J. Ophthalmol.* **90**, 186–190 (2006).
55. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **35**, 24824–24837 (2022).
56. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–157 (1947).

57. Paszke, A. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035 (NeurIPS, 2019).
58. Sutton, J. et al. Developing and validating a multivariable prediction model which predicts progression of intermediate to late age-related macular degeneration-the PINNACLE trial protocol. *Eye* **37**, 1275–1283 (2023).

Acknowledgements

We thank Angela Cree for her administrative support of this work. This research was funded in whole or in part by the Wellcome Trust [210572/Z/18/Z]. For the purpose of open access, the author has applied a CC- BY public copyright licence to any author accepted manuscript version arising from this submission. The project has also been funded by ERC Advanced Grant Deep4MI (884622) and by ERSRC grant EP/Y015665/1.. M.J.M. is funded by the German Research Foundation under project 532139938.

Author contributions

R.H and M.J.M. conceived and designed the study, in addition to writing the original manuscript. T.R.P.T. and C.H. contributed textual annotations, and T.R.P.T., C.H, S.R., J.M., M.P. and D.M. contributed image labels. P.H., P.M., J.C.P. and D.R. provided technical guidance and testing of open source code. H.P.N.S., H.B. and U.S.E. provided feedback on the manuscript. S.S. and A.J.L. performed the reader study and label arbitration. All authors critically reviewed and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01893-8.

Correspondence and requests for materials should be addressed to Robbie Holland.

On behalf of the PINNACLE consortium

Robbie Holland[1] ✉, Thomas R. P. Taylor[2], Christopher Holmes[3], Sophie Riedl[4,5], Julia Mai[4,5], Maria Patsiamanidi[2], Dimitra Mitsopoulou[2], Toby Prevost[3], Lars Fritsche[12], Kristina Pfau[7], Maximilian Pfau[8,13], Hendrik P. N. Scholl[7,8,9], Hrvoje Bogunović[10], Ursula Schmidt-Erfurth[4], Daniel Rueckert[1,6,11,14], Sobha Sivaprasad[3,14], Andrew J. Lotery[2,14] & Martin J. Menten[1,6,11,14]

[12]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA. [13]Department of Ophthalmology, University of Bonn, Bonn, Germany.