A Supplementary material

A.1 Definitions and roles of the junior and senior ophthalmologists

In this study the terms 'junior ophthalmologist' and 'senior ophthalmologists' distinguish between clinicians in non-consultant and consultant roles, respectively. This categorization was used to allocate tasks related to data annotation and arbitration. In particular, six junior ophthalmologists provided first- and second-reader labels on the test dataset for the analysis in Sections 2.3, 2.4 and 2.5. Two of the junior ophthalmologists curated specialist reports used to generate curriculum part 2 (see Section 5.4.2). Finally, two senior ophthalmologists arbitrated the first- and second-reader test labels, and assessed the reports written by both the junior ophthalmologists and the VLMs in Figure 4.

The reported duration of experience ophthalmologists was counted as time spent practicing in ophthalmological clinics, and did include time spent pursuing their medical degree. All ophthalmologists practice at leading hospitals and ophthalmology institutions in the UK and Austria, including Southampton Eye Unit, Moorfields Eye Hospital and the Medical University of Vienna, and are considered experts in the interpretation of retinal OCT images for AMD.

A.2 Data curation flowchart

We summarize the curation and use of all datasets and annotations collected for this study in a flowchart in Figure 17.

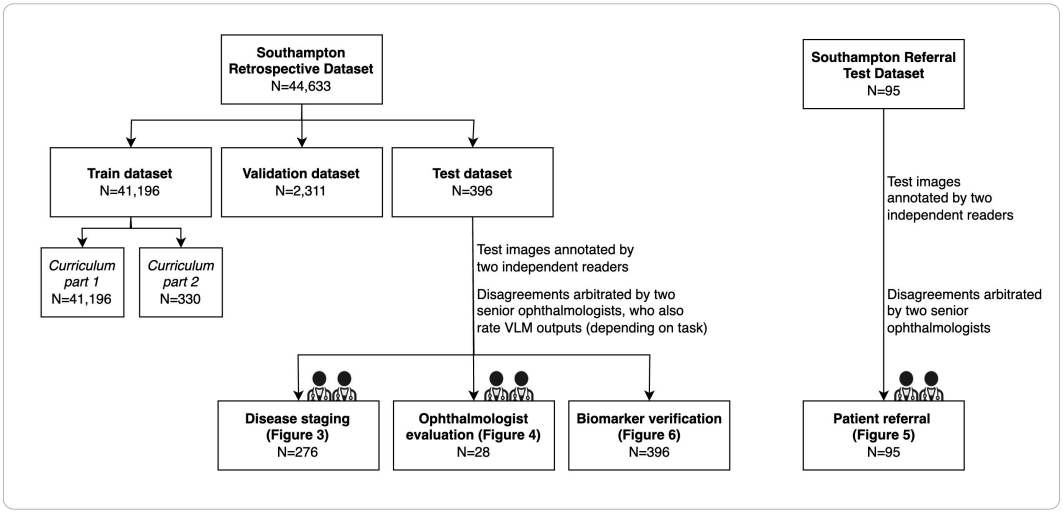

Fig. 17: Flowchart summarizing the curation of data used to train, validate RetinaVLM models, and test data used to evaluate all VLMs and junior ophthalmologists.

A.3 NLP metrics for report generation

Table 1: NLP performance metrics for LLaVA-Med and RetinaVLM-Specialist (rounded to two decimal places).

Model	BLEU-1	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
LLaVA-Med	12.16	1.15	16.33	19.90	4.36	12.88	83.58
RetinaVLM-Specialist	32.11	7.77	29.11	44.16	17.19	29.16	89.49

In addition to the evaluations by senior ophthalmologists in Figure 4, we also evaluate reports generated by VLMs using using a series of natural language processing (NLP) metrics. We have computed seven metrics including BLEU-1 and BLEU-4 [47], METEOR [48], ROUGE-1, ROUGE-2 and ROUGE-L [49], and BERTSCORE-F1 [50]. Each was used to compare the performance of the Llava-Med foundation model and RetinaVLM-Specialist against the junior ophthalmologists.

We find that RetinaVLM-Specialist achieved higher scores than LLaVA-Med in every metric (see Table 1), supporting the senior ophthalmologists' findings of improved performance over Llava-Med. The relative improvement in performance was high in all metrics except for BERTScore, which is known to be insensitive to subtle differences in radiological reports.

A.4 Ablation experiments on hospital and scanner type

Our RetinaVLM models were trained on Topcon OCT images from the Southampton Eye Unit. However, in order to begin to understand the limits of these models we qualitatively test them on three small datasets that contain images from different hospitals and scanners. All images used in this experiment were processed using the same method as those in the main study (see Section 5.1.1 for details). To evaluate VLMs on these images we also used the same prompts as in the clinical evaluation study described in Section 5.7.2.

In the first scenario we tested VLMs on Topcon OCT images from another hospital. On images from Moorfields Eye Hospital we did not find any noticeable degradation in performance compared to images from Southampton Eye Unit (see Figure 18a).

We also test RetinaVLM on scans from the Heidelberg Spectralis scanner. These typically have a higher signal to noise ratio than Topcon scans, though are less widely available. Despite not being trained on Heidelberg Spectralis images, in many cases RetinaVLM-Specialist performs similarly well as on Topcon images (see Figure 18 b). We attribute the relatively strong performance to the use of contrastive learning to train the image encoder. Contrastive learning explicitly trains models to become robust to variations in brightness, contrast, and other visual augmentations, which could lead to better generalization across OCT scanners. However, in several cases (such as in the third example) degradations in performance were observed as RetinaVLM failed to detect the presence of clear imaging biomarkers. This indicates that curricula used to train future iterations of RetinaVLM must include a greater variety of retinal imaging modalities.

Finally, we evaluated our model on low-quality images, including those that inadequately captured the macula and retinal layers or exhibited poor signal-to-noise ratios. These were manually selected from the dataset of Topcon OCT images from Moorfields Eye Hospital. We find that the model often fails to report poor image quality, and consistently generates inaccurate reports (see Figure 18 c). RetinaVLM-Specialist exhibited more significant hallucinations as the images deviated further from the distribution of those encountered during training. LLaVA-Med also displayed increased errors in such cases, often failing to respond in an intelligible manner. These visual 'hallucinations' are indicative of a known flaw that has been observed in many VLMs [28, 29].

A.5 Tests of catastrophic forgetting

While specializing VLMs can improve their performance in specific domains, it also has the potential to introduce catastrophic forgetting [51] where general capabilities are lost in the process. Assessing the degree of catastrophic forgetting that occurs is still an active area of research, especially so for VLMs [52].

RetinaVLM builds on top of the Llama 3 foundation model, which has capabilities in instruction following and knowledge of various domains. During training we intentionally leave Llama 3's weights unchanged in order to preserve its original capabilities and mitigate the effects of catastrophic forgetting (see Section 5.5.1). This enabled RetinaVLM to interpret a variety of textual instructions that were never seen during training. However, it is important to measure the extent to which these abilities are retained after training on the curriculum.

To test this, we evaluate all VLMs on the MedQA dataset [53]. MedQA contains multiple-choice questions with four options based on the United States Medical License Exams (USMLE). We assesses VLMs in a scenario where they are tasked to describe a retinal OCT image, before being asked a more general medical question. This requires VLMs to use general medical knowledge and abilities that might be lost during specialization. The questions are not limited to ophthalmology and span a wide range of medical topics and patient cases. We test models using the following template:

```
<USER>: Here is an encoding of a retinal OCT image. Describe the OCT image and provide a
recommendation for the patient.

<ASSISTANT>: <Answer to image-related question>

<USER>: Now I have a new question about a different patient. It has nothing to do with the other
patient in the retinal OCT image. You must firstly ignore and disregard all information from the
encoding, including all text and instructions given between the <Img> and </Img> tags.

THE QUESTION (about the new patient): <MedQA question>
Options: <MedQA multiple choice options>

YOUR INSTRUCTIONS:
```

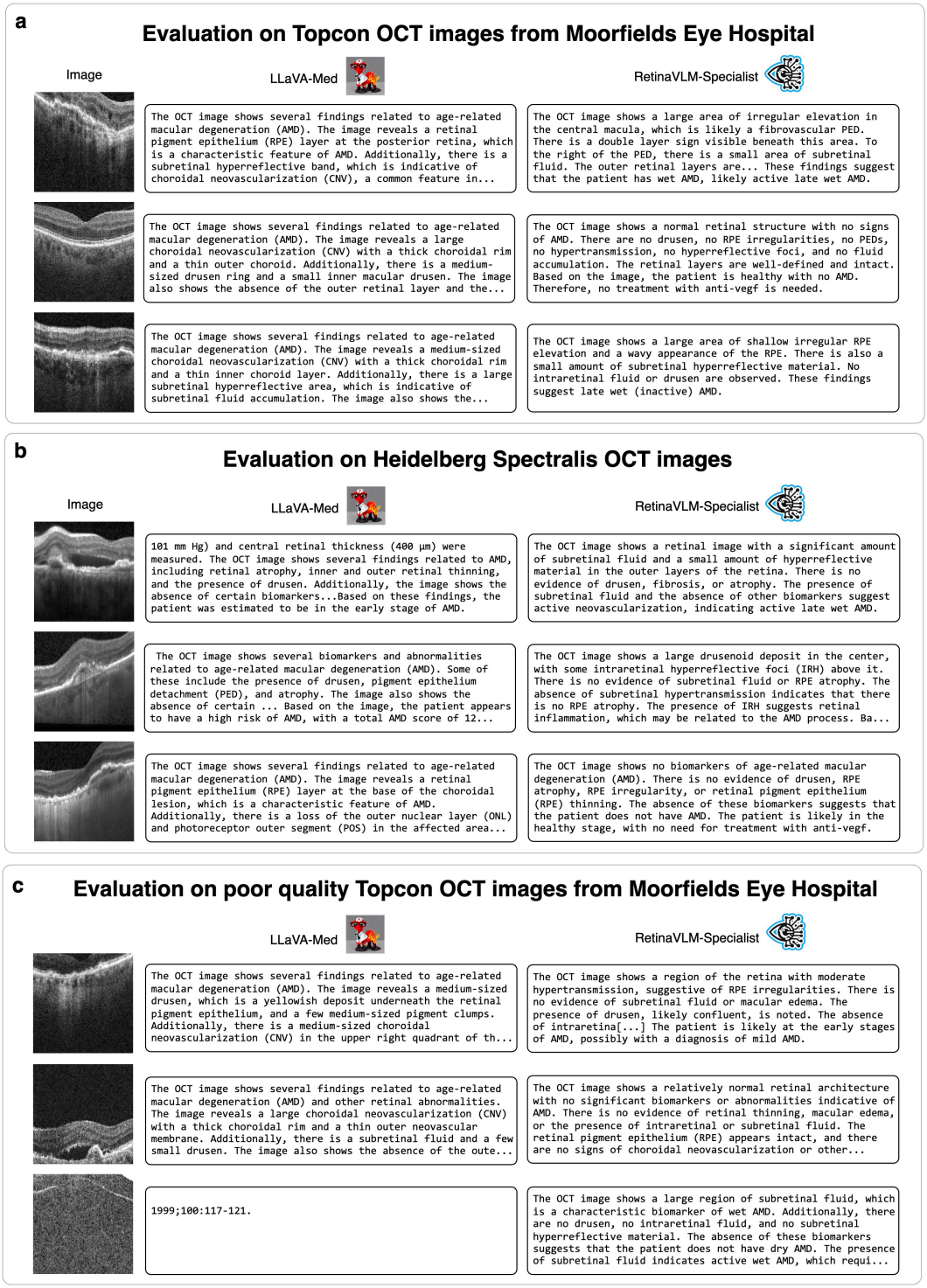

Fig. 18: RetinaVLM-Specialist and LLaVA-Med performance on external datasets and image types. **(a)** Testing on Topcon OCT images from Moorfields Eye Hospital showed no observable performance degradation compared to Southampton Eye Unit images. **(b)** Performance on Heidelberg Spectralis scans, which generally have a higher signal-to-noise ratio, remained comparable to Topcon images, despite the model not being specifically trained on them. **(c)** Poor image quality from manually selected Topcon OCT images led to consistently inaccurate or hallucinated outputs, demonstrating the need for improved handling of low-quality data in future versions of RetinaVLM.

```
1. Do not write about the old patient in the retinal OCT scan.
2. Read the question about the new patient and explain all your thoughts step by step. Write at
great length and in great detail. Do this as if you are an expert in medicine.
3. At the end of your response, conclude by stating the most likely of the four listed options
answers the question. The answer can always be determined.

<ASSISTANT>: <Answer to MedQA question>
```

Similar to [52], we extract the predictions of each VLM and compute their accuracy on the 1279 question-answer pairs in the MedQA test dataset. We also compare their performance against a text-only upper-bound, which is computed by providing the original Llama 3 model with MedQA question alone and no image input. We refer to this baseline as 'Llama 3 (text only)'.

We find that the text-only baseline scores 53.2% accuracy (see Table 2). Compared to this, Med-Flamingo (31.9%) and LLava-Med (18.7%) perform very poorly, scoring around or below random guessing (25%). RetinaVLM-Base and RetinaVLM-Specialist perform better, scoring 52.8% and 44.8%, respectively. The reduction in performance from RetinaVLM-Base to RetinaVLM-Specialist does indicate that the second finetuning step incurs some catastrophic forgetting.

In particular, we observed that RetinaVLM-Specialist output more concise responses than RetinaVLM-Base. This characteristic did not affect its ability to follow complex instructions related to the interpretation of retinal OCT images in the main study. However, on MedQA it reduced its ability to perform multi-step chain-of-thought reasoning, which is known to improve the performance of models on standardized tests [43]. Future future work can address this reduction by increasing the diversity of curriculum part 2 to involve longer, multimodal questions that require multi-step solutions.

Table 2: MedQA accuracy (4 options) for baseline foundation medical VLMs, RetinaVLM models, and a text-only upper bound using Llama 3 8B-Instruct.

Model	**MedQA Accuracy (4 options)**
Med-Flamingo	31.9%
LLaVA-Med	18.7%
RetinaVLM-Base	51.3%
RetinaVLM-Specialist	42.5%
Llama 3 8B-Instruct (text only)	53.2%

A.6 Question-answer generation prompts (curriculum part 1: introduction to retina)

The limited scope of the first part of the tabular reports led us to use only one prompt (see Section A.6.1) to create the 408,545 question-answers in curriculum part 1, 'introduction to retina'.

A.6.1 General question-answering module

I am constructing a dataset to train a model to answer questions based solely on OCT images.

The model will access ONLY the image to deduce attributes. For context, the image in question has attributes as follows:

<ReportText>

Based on these attributes, generate a numbered list of diverse questions and answers.

Ensure the format is:

1. Q: [Question about an image attribute]

A: [Specific answer deduced from the image]

Rules:

- Questions should be crafted in a way that they don't explicitly state the attribute values, but the answers should be based on them. All attributes can be determined from the image.

- Incorporate both yes/no and open-ended styled questions, but always provide a definitive answer in the answer section.

- Occasionally touch on patient outcome/treatment.

A.7 Question-answer generation prompts (curriculum part 2: advanced retinal specialism)

Creating a large quantity and variety of question-answer pairs was important for preventing overfit when finetuning on the 330 images that constitute curriculum part 2. The prompts each focus on a different aspect of question answering and report writing, but are expected to have some overlap in the question-answer pairs they create. The six modules used to generate up to 230 question-answers per image report, as shown in Figure 2, are covered by the following 10 prompts.

1. Advanced biomarkers (Up to 30 total question-answers)
   - Up to 30 question-answer pairs specific to biomarkers (see Section A.7.1)
2. Disease staging definitions (Up to 50 total question-answers)
   - Up to 20 question-answers that use the observation and staging guidelines (see Section A.7.2)
   - Up to 30 question-answers that train the model to replicate any reasoning linking observations to the disease stage that is present in the report (see Section A.7.3)
3. Staging reasoning (Up to 60 total question-answers)
   - Up to 20 question-answers that use the observation and staging guidelines to train the model to explain the diagnosed disease stage in terms of the visible biomarkers (see Section A.7.4)
   - Up to 40 question-answers that focus on training the model in differentiating active from inactive late wet AMD (see Section A.7.5)
4. Referral reasoning (Up to 25 total question-answers)
   - Up to 25 question-answer pairs that focus on training the model to reason about different levels of urgency in patient referral (see Section A.7.6)
5. General visual question-answering (Up to 30 total question-answers)
   - Up to 15 question-answers created directly from the report, but specific to one attribute (see Section A.7.7)
   - Up to 15 question-answers created directly from the report, but including longer and more open-ended questions A.7.8)
6. Report writing (Up to 35 total question-answers)
   - Up to 15 question-answers that focus on replicating the specialist's report (see Section A.7.9)
   - Up to 20 question-answers that all three observational, disease staging and patient referral guidelines to train the model to create more sophisticated reports A.7.10)

A.7.1 Advanced biomarkers module

```
I am constructing a dataset to train a model to answer questions based solely on OCT images.

Below are guidelines outlining what can appear in an OCT image:
<ObservationGuidelines>

However, the image the model is being asked about is characterised by the following description:
DESCRIPTION OF IMAGE: "<ReportText>"

Task: Write 30 varied questions and answers that ask the model about the image.

Ensure the format is:
  1. Q: [Question or statement to describe the image]
  A: [Augmented version of actual image description]

Rules:
```

```
- Ask separately about the presence, amount, location and type of some of the biomarkers in the
guidelines

- Ask about the presence or absence of the biomarkers in the guidelines

- Rather than saying the image/description does not specify/mention a biomarker, instead say 'the
image does not show/display/exhibit evidence of' the biomarker UNLESS its presence is already
implied by another present biomarker. The model does not see the above description of the image,
it's only given the original image when answering questions.

- Try not to give away too much information included in the image description in the question
text. The model must learn to use the image to determine the answer, and not make educated
guesses based on the question alone.

- The answer must be accurate and reflect the same information in the image description.

- Write nothing except the questions and answers.


Tips:

- Example questions: "Does this image show any subretinal fluid?" "Do you see any intraretinal
fluid?" "Is there any hypertransmission? "Does the image show a PED?"

- Answer style variation: Finally, do not start too many questions with "No, ..." or "Yes, ...".
Vary the answer style (the biomarker is 'not present', 'is shown', 'exhibits no', 'does contain'
etc...)

- Positive and negative balance: Try to include an even balance of questions with positive
responses (i.e. ask the model about each of the biomarkers that were reported in the description)
and negative responses (i.e. that biomarker is not present)

  To do this, if you create a question about a biomarker that isn't in the description, try to
  create a second, similar question about a biomarker that is observable in the image.

  In order to make the question set not give too much away about the image, you can make paired
  questions which have positive and negative responses.

  For example, for an image with a PED but no subretinal fluid, if you ask f.e.

    "Q: Is there a PED? If so, where? A: There is a PED present, it's in the center..."

  you should also create a question with a negated answer f.e.

    "Q: Is there any subretinal fluid? And in what quantity? A: There is no sign of subretinal
    fluid in the image...".

  This will help you keep an even balance of positive and negative responses, so that the model
  cannot guess the answer to the question without considering the image.
```

A.7.2 Disease staging definitions modules

```
I am constructing a dataset to train a model to answer questions based solely on OCT images.


Below are guidelines outlining what can appear in an OCT image:

<ObservationGuidelines>

<DiseaseStagingGuidelines>


However, the image the model is being asked about is characterised by the following description:

DESCRIPTION OF IMAGE: "<ReportText>"


Task: Write 20 varied questions and answers that ask the model about the image.


Ensure the format is:

  1. Q: [Question or statement to describe the image]
```

```
  A: [Augmented version of actual image description]


Rules:

- Information about the image should not be in the question.

- The answer must be accurate and reflect the same information in the image description.

- Write nothing except the questions and answers.


Tips:

- Questions should be specific and ask about certain attributes, or sets of attributes. For
example "Q: Is there any subretinal fluid in this image?" or "Q: Is the AMD stage intermediate,
or is it more advanced?".

- Ask separately about the presence, amount, location and type of some of the biomarkers. Try to
create an even balance of 'yes and 'no' answers.

- Rather than saying the image does not directly/explicitly specify/mention the presence of an
attribute, instead say 'the image does not exhibit/show/display/evidence' the attribute, UNLESS
its presence is already directly implied by the presence of another attribute. The model does not
see the above description of the image, it's only given the original image when answering
questions.

- Sometimes the desired output format should be specific in the question (f.e. Answer with 'yes'
or 'no', or answer by stating if the image 'does' or 'does not' contain the biomarker in
question.)
```

A.7.3 Disease staging from the report module

```
I am constructing a dataset to train a model to answer questions based solely on OCT images.


Common disease stages for age-related macular degeneration are: healthy (no-AMD), non-AMD
pathology, early AMD, intermediate AMD, late dry AMD, late wet (inactive) AMD, late wet (active)
AMD

These stages may be referred to with slightly different names in the image description.

If the description doesn't specify whether the late wet AMD is inactive or active, then you
shouldn't either (just see how to description refers to it, but don't feel the need to copy it
verbatim)


However, the image the model is being asked about is characterised by the following description:

DESCRIPTION OF IMAGE: "<ReportText>"


Task: Write 30 varied questions and answers that require the model to estimate the patient's
disease stage from the image.


Ensure the format is enumerated:

  1. Q: [Question or statement to describe the image]

  A: [Augmented version of actual image description]


Rules:

- If the estimated disease stage is not provided in the report, do not write ANY questions and
answers. Simply write "No disease stage in report".

- Try not to give away too much information included in the image description in the question
text. The model must learn to use the image to determine the answer, and not make educated
guesses based on the question alone.
```

```
- The answer must be accurate and reflect the same information in the image description.

- The questions and answers must vary in their style, formulation and vocabulary.

- Write nothing except the questions and answers.


Tips:

- Most questions should ask the model to first estimate the disease stage

  For example, "Decide the most advanced AMD stage supported by the image, and explain your
  reasoning by noting any biomarkers most relevant to that stage."

- If the explanation or reasoning for the disease stage is given in the report, make sure to
include that in the model's answers

- Questions 1 to 20 should first ask the model to decide/estimate/determine/identify/... the
(most advanced) disease stage, and then explain their answer.

- Questions 21 to 30 questions should ask for the biomarkers and then the disease stage, such as
"Describe any relevant/notable/significant biomarkers, and link them to the most likely disease
stage" (which will be the one in the report)
```

A.7.4 Disease staging reasoning (with guidelines)

```
I am constructing a dataset to train a model to answer questions based solely on OCT images.


Below is a schema outlining what can appear in an OCT image:

<ObservationGuidelines>

<DiseaseStagingGuidelines>


However, the image the model is being asked about is characterised by the following description:

DESCRIPTION OF IMAGE: "<ReportText>"


Task: Write 20 varied questions and answers that require the model to perform chain-of-thought
reasoning.


Ensure the format is:

  1. Q: [Question or statement to describe the image]

  A: [Augmented version of actual image description]


Rules:

- Information about the image listed in the description should not be used in the question, only
in the answer.

- Never use the word 'description' or 'mention' in the answer, the model does not see the
description of the image, it only sees the original image itself.

- The answer must be accurate and reflect the same information in the image description.

- Write nothing except the questions and answers.


Tips:
```

```
- Some questions should ask the model to provide a long and detailed answer to questions like
'Describe all the observable biomarkers in the image and then link these to the most likely
disease stage'.

- One or two questions should ask the model to explain/deduce the highest precedent AMD stage
(i.e. the overall AMD stage) by describing the biomarker(s) in the image which belong to the most
advanced disease stage, and linking them to the relevant disease stage.

- Some questions should ask the model to summarise any relevant biomarkers and, explaining its
reasoning using the guidelines, conclude with the AMD stage. For example, a drusenoid PED
suggests intermediate AMD, but if coupled with subretinal fluid the overall AMD stage becomes
active late wet due to the fluid.

- Some questions should ask the model to fully describe and list all its image observations, and
then conclude the presence, absence, location or quantity of a specific biomarker (f.e. 'Describe
the OCT image in detail and note any abnormalities, and then tell me if the image contains
subretinal fluid.')
```

A.7.5 Disease staging reasoning second module

```
 I am constructing a dataset to train a model to answer questions based solely on OCT images.


Below are guidelines outlining what can appear in an OCT image:

<DiseaseStagingGuidelines>


However, the image the model is being asked about is characterised by the following description:

DESCRIPTION OF IMAGE: "<ReportText>"


Task: Write 40 varied questions and answers that require the model to estimate the patient's
disease stage from the image.


Ensure the format is enumerated:

  1. Q: [Question or statement to describe the image]

  A: [Augmented version of actual image description]


Rules that always apply:

- If the estimated disease stage is not provided in the report, do not write ANY questions and
answers. Simply write "No disease stage in report".

- If reasoning for the disease stage is provided in the report, you MUST include this
logic/nuance in the model's answers.

- You MUST not give away information about the image description in the question text. The model
must learn to use the image to determine the answer, and not make educated guesses based on the
question alone.

- The answer must be accurate and reflect the same information in the image description.

- The questions and answers must vary in their style, formulation and vocabulary.

- Write nothing except the questions and answers.


Rules that are specific to differentiating cases of active, vs inactive, late wet AMD:

- In cases with an active late wet diagnosis, you MUST make it clear that the PRESENCE OF FLUID
is what differentiates active from inactive late wet AMD. Make this clear when explaining the
reasoning behind active late wet diagnoses.

  So do not imply that f.e. "The disease stage is late wet AMD (active), as indicated by the
  subretinal/intraretinal fluid, subretinal hyperreflective material, ...."
```

```
  Instead, you MUST explain that f.e. "The overall disease stage is late wet AMD, which is active
  due to the detection/presence of subretinal/intraretinal fluid. Inactive late wet features
  include fibrovascular PED, ..."

  Or "Late wet AMD best describes the AMD stage. The detection/presence of
  subretinal/intraretinal fluid means this is active late wet AMD. Other late wet features
  include fibrovascular PED, ..."

- Similarly, in cases with an inactive late wet diagnosis, you MUST make it clear in the model's
reasoning that the lack of fluid, in combination with the other biomarkers, is what resulted in
the inactive late wet diagnosis.

  For example, "The stage is late wet AMD according to the evidence/presence of subretinal
  hyperreflective material, but it is inactive as there is no detectable fluid of any kind in the
  image"

- The exact formulation of this answer must vary according to the question. Do not copy the
examples too many times. Add a lot of diversity in the model's responses.


Tips:

- Most questions should ask the model to first estimate the disease stage

- Some questions should first ask the model for its disease stage prediction, and then ask it to
explain its reasoning by noting visible biomarkers that relate to that stage

  For example, "Decide the most advanced AMD stage supported by the image, and explain your
  reasoning by noting any biomarkers most relevant to that stage."

- The explanation for the disease stage may already be given in the report, but you can also use
the guidelines provided to work out which biomarkers resulted in that disease stage.

- Questions 31 to 40 questions should ask for the biomarkers and then the disease stage, such as
"Describe any relevant/notable/significant biomarkers, and link them to the most likely disease
stage" (which will be the one in the report)
```

A.7.6 Patient referral reasoning module

```
I am constructing a dataset to train a model to answer questions based solely on OCT images.


Below are guidelines outlining what can appear in an OCT image:

<DiseaseStagingGuidelines>


<PatientReferralGuidelines>


However, the image the model is being asked about is characterised by the following description:

DESCRIPTION OF IMAGE: "<ReportText>"


Task: Write 25 varied questions and answers that require the model to perform explain its
reasoning before making conclusions and recommendations.


Ensure the format is:

  1. Q: [Question or statement to describe the image]

  A: [Augmented version of actual image description]


Rules:

- Information about the image should not be in the question.

- Never use the word 'description' or 'mention' in the answer, the model does not see the
description of the image, it only sees the original image itself.
```

- The answer must be accurate and reflect the same information in the image description.

- Write nothing except the questions and answers.

- The model sees each questions separately so they will not be seen together.

Tips:

- Some questions should ask the model to list any relevant biomarkers and, based on these, recommend if the patient should be referred or not.

- Many questions should make a series of requests, by asking the model to write a long and detailed answer reporting all the observable biomarkers, linking those to the most likely disease stage and then summarising the report with a patient referral recommendation.

  For example, 'Describe the all biomarkers in the image, and based off of your observations which AMD best describe the patient. Summarise your report with a referral recommendation that follows the treamtent guidelines.'

- Some questions should ask the model to recommend if the patient does: not need referral, if they need general attention by a specialist, or if they likely need treatment with anti-vegf based off the models observations

- Some questions should ask the model to explain/deduce/estimate the patient's risk (i.e. the recommended referral action) by first describing the most advanced or concerning biomarker(s) (i.e. the observable biomarker(s) which belong to the most severe disease stage)

A.7.7 Specific report-based questions module

I am constructing a dataset to train a model to answer questions based solely on OCT images.

The image in question is characterised by the following description:

DESCRIPTION OF IMAGE: "<ReportText>"

Task: Write 15 varied questions and answers that ask the model about the image. They should

Ensure the format is:

1. Q: [Question or statement to describe the image]

   A: [Augmented version of actual image description]

Rules:

- Try not to give away too much information included in the image description in the question text. The model must learn to use the image to determine the answer, and not make educated guesses based on the question alone.

- The answer must be accurate and reflect the same information in the image description.

- Write nothing except the questions and answers.

Tips:

- Questions should be specific and ask about certain attributes, or sets of attributes. For example "Q: Is there any subretinal fluid in this image?".

A.7.8 General report-based questions module

I am constructing a dataset to train a model to answer questions based solely on OCT images.

The image in question is characterised by the following description:

```
DESCRIPTION OF IMAGE: "<ReportText>"


Task: Write 15 varied questions and answers that ask the model about the image.


Ensure the format is:

  1. Q: [Question or statement to describe the image]

  A: [Augmented version of actual image description]


Rules:

- The answer should be a modified and augmented version of the actual description.

- Try not to give away too much information included in the image description in the question
text. The model must learn to use the image to determine the answer, and not make educated
guesses based on the question alone.

- The answer must be accurate and contain the same information as the actual description.
However, the order of the sentences as they are written must change and randomly vary.

- Write nothing except the questions and answers.


Tips:

- Some question should be general and ask to describe the image.

- Other questions should be more specific such as "Q: Is there any subretinal fluid in this
image?" that have shorter answers.

- Example questions/statements might be "Describe the OCT image in detail." or "Can you give me a
summary of the image?".
```

A.7.9 Basic report writing module

```
I am constructing a dataset to train a model to answer questions based solely on OCT images.


The image in question is characterised by the following description.

DESCRIPTION OF IMAGE: "<ReportText>"


Task: Write 15 questions and answers that ask the model to describe the image in full. The first
ten questions should ask to describe the entire image (with jumbled/permuted/randomised answers),
while the final five should be more specific or use segments of the description.


Ensure the format is:

  1. Q: [Question or statement to describe the image]

  A: [Augmented version of actual image description]


Rules:

- The answer should be a modified and augmented version of the actual description.

- The answer should contain the same information as the actual description but the order of the
sentences as they are written must change and randomly vary

- The question must not contain any information about the image.

- Write nothing except the questions and answers.
```

```
Tips:

- Example questions/statements might be "Describe the OCT image in detail." or "Can you give me a
summary of the image?" or "Write a report on this image to be given to an optometrist.".
```

A.7.10 Advanced report writing module

```
I am constructing a dataset to train a model to answer questions based solely on OCT images.


Below are guidelines outlining what can appear in an OCT image:

<ObservationGuidelines>

<DiseaseStagingGuidelines>

<PatientReferralGuidelines>


However, the image the model is being asked about is characterised by the following description:

DESCRIPTION OF IMAGE: "<ReportText>"


Task: Write 20 requests for reports that ask the model reason all the way from the observable
biomarkers, linking these to the most advanced disease stage, and finally to a referral
recommendation for the patient.


Ensure the format is:

  1. Q: [Request to the model]

  A: [Answer or report deduced using actual image description]


Rules:

- Never use the word 'description' or 'mention' in the answer, the model does not see the
description of the image, it only sees the original image itself.

- Try not to give away too much information included in the image description in the question
text. The model must learn to use the image to determine the answer, and not make educated
guesses based on the question alone.

- Write nothing except the questions and answers.

- The model sees each questions separately so they will not be seen together.


Tips:

- Example questions/statements might be "Describe the OCT image in detail." or "Can you give me a
summary of the image?" or "Write a report on this image to be given to a retinal specialist.".

- Ask the model to explain why the AMD stage is not more advanced than it is (i.e. the absence of
certain late stage biomarkers)

- Ask the model to explain why the AMD stage is more advanced than an earlier stage (i.e. the
presence of certain late stage biomarkers)

- A few times, request verbose reports like "Write a report that starts by highlighting the most
significant and salient biomarkers, and link these to the most probable disease stage. Conclude
with a referral recommendation."
```