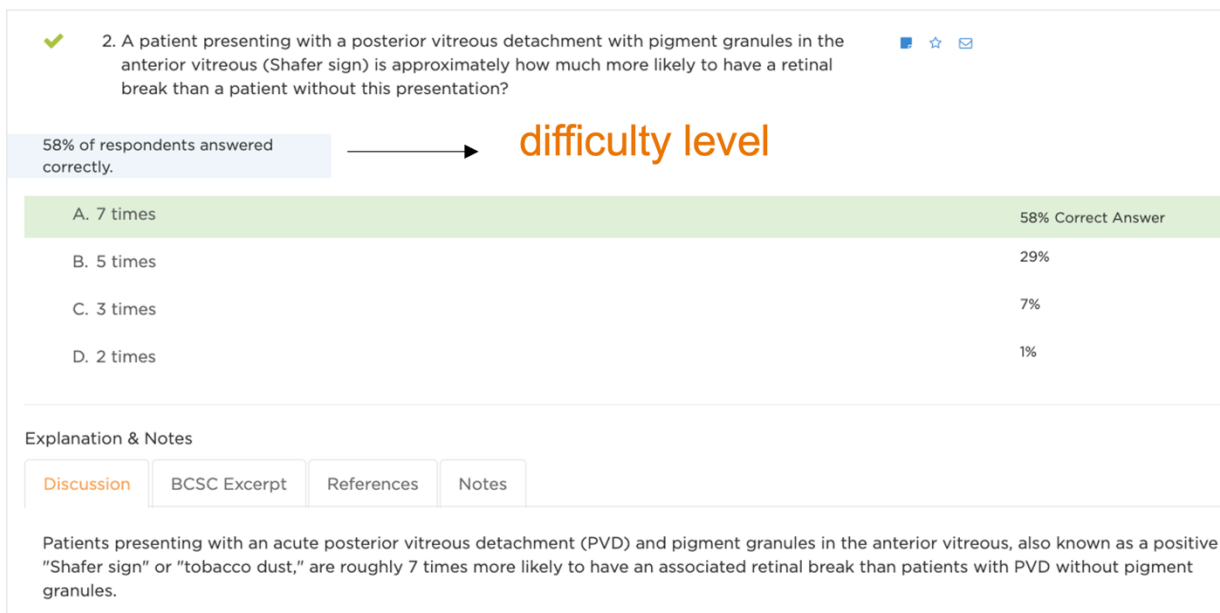


Supplementary Materials

1. Data Acquisition and Description

BCSC Self-Assessment Program and OphthoQuestions question bank are widely used by ophthalmology residents, fellows, and practicing ophthalmologists in the United States and Canada to prepare for essential exams such as the Ophthalmic Knowledge Assessment Program (OKAP) and the American Board of Ophthalmology (ABO) qualifying exam. They are both online platforms; their web-based application provides interactive questions that align with the latest clinical knowledge and offer detailed feedback and relevant resources (See Supplementary Figure 1). As of January 2023, BCSC has 4,458 questions and OphthoQuestions 4,539 [13,26].



2. A patient presenting with a posterior vitreous detachment with pigment granules in the anterior vitreous (Shafer sign) is approximately how much more likely to have a retinal break than a patient without this presentation?

58% of respondents answered correctly. → difficulty level

A. 7 times	58% Correct Answer
B. 5 times	29%
C. 3 times	7%
D. 2 times	1%

Explanation & Notes

Discussion | BCSC Excerpt | References | Notes

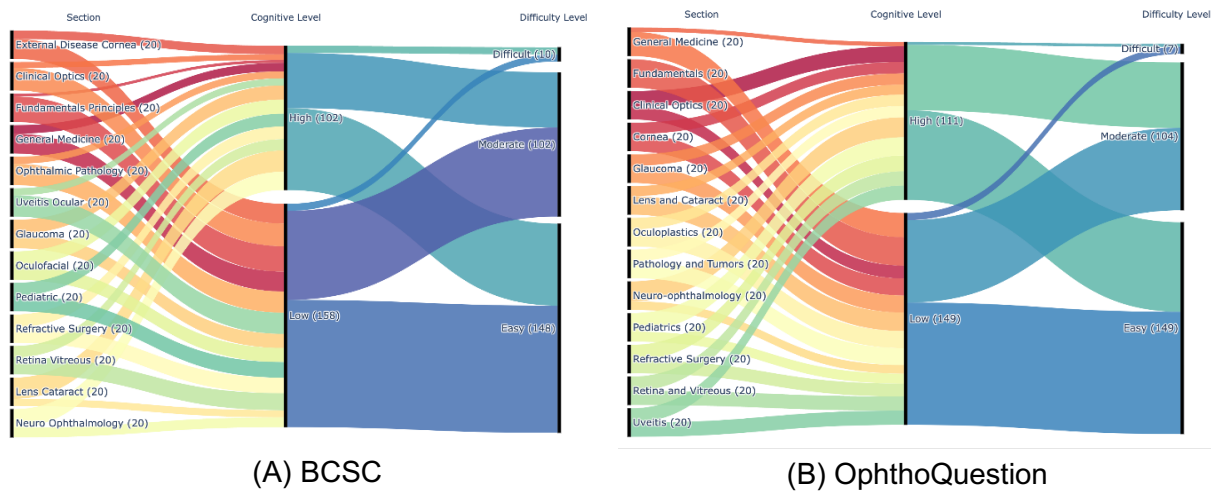
Patients presenting with an acute posterior vitreous detachment (PVD) and pigment granules in the anterior vitreous, also known as a positive "Shafer sign" or "tobacco dust," are roughly 7 times more likely to have an associated retinal break than patients with PVD without pigment granules.

Supplementary Figure 1: Example of a clinical question from the BCSC Self-Assessment platform. Each question includes the question text, four answer options, a highlighted correct answer (when the choice has been given) and percentage-based respondent data which can be used as an indicator of the question's difficulty level. Below the question, additional sections provide a discussion, references, and explanatory notes to support learning. OphthoQuestions question has the same structure.

Difficulty Level can be obtained straightforwardly by classifying the percentage of correct responses into Difficult (<30%), Moderate ($\geq 30\%$ and <70%), and Easy ($\geq 70\%$) [13,30]. Cognitive Level are defined in the OKAP User's Guide [31] as followed:

- **Level 1 – Recall:** assesses an examinee's understanding of core facts, concepts, and procedures in ophthalmology, requiring more than simple memorization. They test accuracy, comprehension, and the ability to distinguish subtle differences and recognize procedural errors.
- **Level 2 – Interpretation:** involves recalling facts and interpreting their implications. Examinees must make inferences, predict outcomes, and identify abnormalities, often in unfamiliar scenarios, requiring more than just recall.
- **Level 3 – Decision Making & Clinical Management:** tests the application of knowledge to complex situations, such as making differential diagnoses and managing cases. These items require problem-solving in unique circumstances where prior instruction may not be directly applicable.

However, as explained in [13], level 2 and 3 can be challenging to distinguishing at times. Thus, only two level Low (Level 1) and High (Level 2 and 3) are considered. Low cognitive level questions primarily focus on the recall of facts and concepts, such as identifying the gene associated with a specific condition. In contrast, high cognitive level questions are designed to assess more complex tasks, including data interpretation, performing calculations, and patient management. The distribution of the questions with respect to Cognitive Level and Difficulty Level are illustrated in Supplementary Figure 2.



Supplementary Figure 2: 260 questions sampled from BCSC Self-Assessment Program (A) and OphthoQuestions question bank (B).

2. Q4 vs Q8

Supplementary Table 1: Performance comparison (in % accuracy) between Q8 and Q4 of Llama-3-70B, Gemma-2-27B, and Mixtral-8x7B on BCSC and OphthoQuestions question sets.

Models	BCSC	OphthoQuestions	Mean
--------	------	-----------------	------

		ZRS	ZRS-CoT	RAG	ZRS	ZRS-CoT	RAG	ZRS	ZRS-CoT	RAG
Llama-3-70B	Q4	64.62	70.77	83.46	50.38	65.77	77.31	57.5	68.27	80.385
	Q8	76.54	71.15	84.62	70.38	63.08	76.15	73.46	67.115	80.385
Gemma-2-27B	Q4	64.23	61.54	83.46	60.0	56.92	75.0	62.115	59.23	79.23
	Q8	65.38	63.08	85.0	60.0	56.92	76.92	62.69	60.0	80.96
Mixtral-8x7B	Q4	57.69	56.92	54.62	48.08	52.69	71.92	60.0	58.85	78.46
	Q8	60.0	58.85	78.46	56.92	54.62	70.38	58.46	56.735	74.42

3. Performance by section

Supplementary Table 2a: Performances (%) of GPT-4 and Mixtral-8x7B between zero-shot setting and RAG on BCSC. An additional column which is the difference between RAG and zero-shot is provided to highlight the improvement gain achieved by RAG.

BCSC												
Section@	GPT-4-turbo			Mixtral8-x7B-Q4			Llama-3-70B			Gemma-2-27B		
	ZRS	RAG	Gain	ZRS	RAG	Gain	ZRS	RAG	Gain	ZRS	RAG	Gain
General Medicine	90%	95%	+5%	70%	80%	+10%	65%	95%	+30%	75%	90%	+15%
Fundamentals	70%	100%	+30%	40%	80%	+40%	60%	95%	+35%	55%	95%	+40%
Clinical Optics	80%	95%	+15%	60%	90%	+30%	65%	95%	+30%	70%	90%	+20%
Pathology and Tumors	80%	95%	+15%	60%	85%	+25%	65%	80%	+15%	65%	85%	+20%
Neuro-ophthalmology	80%	100%	+20%	60%	85%	+25%	80%	90%	+10%	80%	85%	+5%
Pediatrics	80%	90%	+10%	60%	65%	+5%	45%	70%	+25%	70%	70%	+0%
Oculoplastics	90%	90%	+0%	65%	70%	+5%	50%	70%	+20%	70%	65%	-5%
Cornea	80%	70%	-10%	60%	75%	+15%	60%	70%	+10%	60%	80%	+20%
Uveitis	85%	100%	+15%	70%	85%	+15%	65%	95%	+30%	65%	100%	+35%
Glaucoma	75%	85%	+10%	45%	70%	+25%	90%	70%	-20%	40%	70%	+30%
Lens and Cataract	75%	90%	+15%	50%	80%	+30%	65%	75%	+10%	60%	85%	+25%
Retina and Vitreous	80%	90%	+10%	50%	75%	+25%	75%	90%	+15%	65%	85%	+20%
Refractive Surgery	80%	95%	+15%	60%	80%	+20%	55%	90%	+35%	60%	85%	+25%
Mean	80.38%	91.92%	+11.54%	57.69%	78.46%	+20.77%	65%	83%	+19%	64%	83%	+19%

Supplementary Table 3b: Performances (%) of GPT-4 and Mixtral-8x7B between zero-shot setting and RAG on OphthoQuestions. An additional column which is the difference between RAG and zero-shot is provided to highlight the improvement gain achieved by RAG.

OphthoQuestions												
Section	GPT-4-turbo			Mixtral8-x7B-Q4			Llama-3-70B			Gemma-2-27B		
	ZRS	RAG	Gain	ZRS	RAG	Gain	ZRS	RAG	Gain	ZRS	RAG	Gain
General Medicine	90%	95%	+5%	60%	75%	+15%	70%	80%	+10%	80%	95%	+15%
Fundamentals	85%	75%	-10%	45%	70%	+25%	65%	75%	+10%	65%	75%	+10%
Clinical Optics	55%	65%	+10%	45%	75%	+30%	20%	70%	+50%	55%	60%	+5%
Pathology and Tumors	65%	75%	+10%	45%	55%	+10%	30%	70%	+40%	55%	70%	+15%
Neuro-ophthalmology	80%	90%	+10%	55%	85%	+30%	75%	75%	+0%	45%	80%	+35%
Pediatrics	75%	90%	+15%	35%	75%	+40%	35%	80%	+45%	60%	65%	+5%
Oculoplastics	85%	95%	+10%	45%	65%	+20%	55%	90%	+35%	60%	80%	+20%
Cornea	75%	75%	+0%	40%	65%	+25%	50%	60%	+10%	55%	70%	+15%
Uveitis	80%	90%	+10%	30%	75%	+45%	25%	65%	+40%	55%	60%	+5%
Glaucoma	65%	80%	+15%	45%	65%	+20%	60%	70%	+10%	65%	75%	+10%
Lens and Cataract	85%	95%	+10%	70%	80%	+10%	75%	85%	+10%	75%	75%	+0%
Retina and Vitreous	95%	90%	-5%	50%	75%	+25%	45%	95%	+50%	60%	85%	+25%
Refractive Surgery	75%	95%	+20%	60%	75%	+15%	50%	90%	+40%	50%	85%	+35%
Mean	77.69%	85.38%	+7.69%	48.08%	71.92%	+23.85%	50%	77%	+27%	60%	75%	+15%

4. Performance by section

