

Supplementary Information

Supplementary Table 1. Inter-rater Reliability Statistics	2
Supplementary Table 2. Model comparison of MedHistory Performance on a Shared Set of Real-world Cases	2
Supplementary Table 3. Similarity Evaluation between Model-generated and Resident-authored Clinical Charts	3
Supplementary Table 4. Ablation Analyses of the Study Agent.....	3
Supplementary Table 5. Item-Level Components and Score Allocation of the MedHistory Framework	4
Supplementary Table 6. Patient Comments from the Satisfaction Questionnaire in the Randomized Clinical Trial.....	5
Supplementary Table 7. Micro-level Quantitative Comparison of Test Recommendation and Primary Diagnosis by Assessment Stage	6
Supplementary Table 8. Comparison of Quality Scores for Secondary and Exploratory Outcomes by Assessment Stage.....	7
Supplementary Table 9. Linear Mixed Effects Estimate on Test Recommendation and Diagnostic Performance	8
Supplementary Table 10. Comparison of Test Recommendation and Diagnostic Performance between high and low Medhistory quality cases	9
Supplementary Table 11. Comparison of Cost and Waiting Time between Groups	10
Supplementary Note 1 General Structure of Symptom-Oriented Logic Trees for Clinical History-Taking	11
Supplementary Note 2 Showcases	15
Supplementary Note 3 Speech-to-text (STT) artifacts.....	33
Supplementary Figure 1. Architecture of the Dual-Agent System with Symptom-Oriented Reasoning	34
Supplementary Note 4 Large Language Model Specifications and Inference Parameters	35
Supplementary Table 12. MedHistory Framework for Ophthalmic History-taking and Documentation .	37
Supplementary Note 5 Prompt Incorporated Ocular Signs in the Randomized Clinical Trial	39
Supplementary Table 13. Qualitative Evaluation Metrics.....	41
Supplementary References.....	41
The CONSORT Reporting Checklist.....	43
Trial Protocol.....	48
Statistical Analysis Plan.....	78

Supplementary Table 1. Inter-rater Reliability Statistics

Theme	Aspect	Intraclass Correlation
Proof-of-Concept Benchmarking (n=20 cases)	Model comparison of MedHistory performance	0.90*
	Similarity evaluation between model and resident charts	0.96 [#]
Randomized Clinical Trial (n=172 participants)	Comparison of MedHistory performance	0.92*
	Quality of test recommendation	0.84 [#]

*Assessed using ICC (2,1) for continuous variables.

[#]Assessed using weighted Cohen's κ for categorical variables.

Supplementary Table 2. Model comparison of MedHistory Performance on a Shared Set of Real-world Cases

MedHistory (Score Range)	Model Name	Model Score	P value
Total (0 to 100)	Study Agent	83.50 (79.38 to 84.00)	Not applicable
	Qwen-2.5 72B	61.00 (51.25 to 64.50)	<0.001
	GPT-4o	60.00 (58.00 to 64.50)	<0.001
	GPT-4.5	64.00 (59.00 to 68.88)	<0.001
	DeepSeek-R1 671B	35.00 (31.50 to 39.38)	<0.001
History-taking (0 to 92)	Study Agent	76.75 (72.62 to 77.00)	Not applicable
	Qwen-2.5 72B	54.00 (44.88 to 58.62)	<0.001
	GPT-4o	54.75 (52.00 to 59.00)	<0.001
	GPT-4.5	60.25 (54.38 to 63.88)	<0.001
	DeepSeek-R1 671B	30.50 (27.38 to 35.62)	<0.001
Communication (0 to 4)	Study Agent	4.00 (4.00 to 4.00)	Not applicable
	Qwen-2.5 72B	4.00 (4.00 to 4.00)	0.09
	GPT-4o	3.00 (3.00 to 3.00)	<0.001
	GPT-4.5	2.00 (2.00 to 2.00)	<0.001
	DeepSeek-R1 671B	3.00 (3.00 to 4.00)	0.002
Documentation (0 to 4)	Study Agent	3.50 (3.00 to 4.00)	Not applicable
	Qwen-2.5 72B	2.50 (2.00 to 3.00)	<0.001
	GPT-4o	3.00 (1.50 to 3.00)	0.04
	GPT-4.5	3.00 (1.50 to 3.00)	0.04
	DeepSeek-R1 671B	1.50 (1.50 to 1.50)	<0.001

Data are median (IQR).

Statistical comparisons between the agent and each of the four LLMs were performed using Wilcoxon signed-rank test (n=20 cases). Qwen-2.5 72B served as the baseline reference.

Supplementary Table 3. Similarity Evaluation between Model-generated and Resident-authored Clinical Charts

Model Name	Similarity score (%)					Median (IQR)
	1	2	3	4	5	
Agent	0	0	1(5)	7(35)	12(60)	5.0 (4.0 to 5.0)
Qwen-2.5 72B	0	6(30)	10(50)	4(20)	0	3.0 (2.0 to 3.0)
GPT-4o	0	2(10)	12(60)	6(30)	0	3.0 (3.0 to 4.0)
GPT-4.5	0	1(5)	9(45)	7(35)	3(15)	3.5 (3.0 to 4.0)
DeepSeek-R1 671B	14(70)	5(25)	1(5)	0	0	1.0 (1.0 to 2.0)

Data are number (%) or median (IQR).

Similarity scores differed significantly across models based on the Friedman rank-sum test ($P < 0.001$; $n = 20$ cases). 1, $< 40\%$ match; 2, 40 to 59% match with major omissions; 3, 60 to 79% match with some unclear key information; 4, $\geq 80\%$ match with minor phrasing issues; 5, $\geq 80\%$ match with all key information clear and complete.

Supplementary Table 4. Ablation Analyses of the Study Agent

MedHistory (Score Range)	Full agent	Without few-shot in feature extraction	Without document optimization module	P value
Documentation (0 to 4)	3.50 (3.00 to 4.00)	2.50 (2.00 to 3.00)	2.00 (2.00 to 2.00)	< 0.001

Data are median (IQR).

The full agent was compared with each of the two ablated conditions using the Wilcoxon signed-rank test ($n = 20$ cases).

Supplementary Table 5. Item-Level Components and Score Allocation of the MedHistory Framework

Content Score (Points)	Agent group	Resident group
Composite score (100)	80.73±6.62	56.94±8.19
History-taking (92)	73.49±6.49	49.93±8.03
Chief Complaint (30)	27.79±3.80	29.71±1.60
Precipitating factor (10)	10.00±0	2.27±4.18
Characteristics of presenting complaint and associated symptoms (10)	9.74±0.86	5.64±1.49
Time of Clinical Course (1)	0.95±0.21	0.17±0.38
Diagnosis of Clinical Course (1)	0.84±0.37	0.15±0.36
Treatment and response of Clinical Course (3)	2.56±1.05	0.48±1.04
Recent progression and evolution (5)	2.48±2.49	0.58±1.61
History of spectacle or contact lens wear (1)	0.98±0.21	0.12±0.32
General health status (4)	3.23±0.79	2.02±0.92
Trauma History (5)	4.91±0.60	2.91±2.48
Ocular Surgical History (5)	4.36±1.34	1.15±2.07
Family History (5)	0.67±1.30	0.20±0.96
Allergy History (5)	2.35±1.23	2.47±0.61
Personal History (5)	0.70±1.31	0.12±0.65
Logic Sequencing (2)	1.93±0.26	1.94±0.32
Communication (4)	3.81±0.47	3.47±0.57
Communication Empathy (2)	1.93±0.26	1.94±0.28
Communication Clarity (2)	1.88±0.32	1.53±0.50
Documentation (4)	3.42±0.58	3.55±0.66
Correct use of medical terminology (2)	1.43±0.52	1.98±0.20
Consistency of dialogue (2)	1.99±0.20	1.57±0.60

Data are mean ± SD.

History-taking (92 points), including chief complaint, history of present illness, past medical history, family history, allergy history, personal history and logic sequencing.

Communication (4 points), rating the interviewer's empathy and clarity of expression.

Documentation (4 points), judging terminology accuracy and dialogue-chart consistency.

Supplementary Table 6. Patient Comments from the Satisfaction Questionnaire in the Randomized Clinical Trial

	Agent group	Resident group
Reasons for dissatisfaction with the AI system	-There was a noticeable delay before the subtitles showed up.	Not applicable
Reasons for reluctance to replace doctors with AI consultation	- AI can't confirm my diagnosis.	-I don't understand AI. -It's not as good as face-to-face communication.
Reasons for being willing to use AI consultation	-I think it's acceptable, since the AI consultation system helps doctors understand the patient's condition in advance. - It was easy to use, and I prefer speaking to typing. - I felt more relaxed than when talking to a doctor. - It was more convenient compared with seeing doctor in the hospital. -It asks questions just like a doctor would.	-It is more convenient for my schedule.

Supplementary Table 7. Micro-level Quantitative Comparison of Test Recommendation and Primary Diagnosis by Assessment Stage

Stage	Outcome	Metric	Agent group (95% CI)	Resident group (95% CI)	Difference Δ (Agent - Resident, 95% CI)	P value
History-only	Test recommendation	Recall	0.62 (0.56 to 0.68)	0.62 (0.58 to 0.67)	-0.01 (-0.08 to 0.07)	0.88
		Precision	0.96 (0.92 to 0.97)	0.97 (0.95 to 0.99)	-0.01 (-0.05 to 0.03)	0.64
		F1	0.75 (0.71 to 0.79)	0.76 (0.73 to 0.79)	-0.01 (-0.06 to 0.05)	0.82
	Primary diagnosis	Recall	0.78 (0.70 to 0.85)	0.56 (0.46 to 0.66)	0.22 (0.09 to 0.33)	<0.001
		Precision	0.94 (0.89 to 0.99)	0.86 (0.79 to 0.94)	0.08 (-0.01 to 0.17)	0.10
		F1	0.85 (0.79 to 0.90)	0.68 (0.59 to 0.76)	0.17 (0.07 to 0.27)	<0.001
Signs-integration	Test recommendation	Recall	0.76 (0.71 to 0.82)	0.74 (0.70 to 0.79)	0.02 (-0.06 to 0.09)	0.62
		Precision	0.96 (0.93 to 0.98)	0.98 (0.96 to 1.00)	-0.02 (-0.05 to 0.01)	0.18
		F1	0.85 (0.81 to 0.89)	0.85 (0.82 to 0.88)	0.00 (-0.05 to 0.05)	0.88
	Primary diagnosis	Recall	0.85 (0.79 to 0.92)	0.81 (0.75 to 0.89)	0.04 (-0.06 to 0.14)	0.41
		Precision	0.90 (0.84 to 0.95)	0.96 (0.92 to 0.99)	-0.06 (-0.13 to 0.01)	0.07
		F1	0.88 (0.83 to 0.92)	0.88 (0.83 to 0.93)	-0.01 (-0.07 to 0.07)	0.85

Data are micro-level metrics across all cases in each group (n=172; 86 per group).

Bootstrap resampling (B = 2000) was used to derive 95% confidence intervals and two-sided P values.

Supplementary Table 8. Comparison of Quality Scores for Secondary and Exploratory Outcomes by Assessment Stage

Stage	Outcome	Agent group (%)				Median (IQR)	Resident group (%)				Median (IQR)	P value
		0	1	2	3		0	1	2	3		
History-only	Test recommendation	0	9 (10.5)	34 (39.5)	43 (50.0)	2.5 (2 to 3)	0	0	62 (72.1)	24 (27.9)	2 (2 to 3)	0.06
	Primary diagnosis	0	4 (4.7)	23 (26.7)	59 (68.6)	3 (2 to 3)	0	4 (4.7)	43 (50.0)	39 (45.3)	2 (2 to 3)	0.004
Signs-integration	Test recommendation	0	5 (5.8)	47 (54.7)	34 (39.5)	2 (2 to 3)	0	0	43 (50.0)	43 (50.0)	2.5 (2 to 3)	0.08
	Primary diagnosis	0	6 (7.0)	33 (38.4)	47 (54.7)	3 (2 to 3)	0	3 (3.5)	25 (29.1)	58 (67.4)	3 (2 to 3)	0.08
	Treatment recommendation	0	6 (7.0)	28 (32.6)	52 (60.5)	3 (2 to 3)	0	2 (2.3)	36 (41.9)	48 (55.8)	3 (2 to 3)	0.75

Data are number (%) or median (IQR).

Group differences were assessed using the Mann–Whitney U test (n=172; 86 per group).

Supplementary Table 9. Linear Mixed Effects Estimate on Test Recommendation and Diagnostic Performance

Outcome	Effect	Δ Recall (95% CI, P value)	Δ Precision (95% CI, P value)	Δ F1 score (95% CI, P value)
Test recommendation	Group* (Agent vs Resident)	0.03 (-0.05 to 0.10; 0.45)	-0.03 (-0.08 to 0.02; 0.21)	0.00 (-0.06 to 0.05; 0.99)
	Stage# (Signs-integration vs History-only)	0.14 (0.08 to 0.20; <0.001)	0.01 (-0.03 to 0.05; 0.65)	0.10 (0.05 to 0.15; <0.001)
	Group $\times$ Stage ⁺	-0.02 (-0.10 to 0.06; 0.67)	0.00 (-0.06 to 0.06; 0.95)	-0.01 (-0.08 to 0.06; 0.83)
Primary diagnosis	Group* (Agent vs Resident)	0.20 (0.09 to 0.32; 0.001)	0.17 (0.06 to 0.28; 0.001)	0.18 (0.08 to 0.30; 0.001)
	Stage# (Signs-integration vs History-only)	0.24 (0.15 to 0.33; <0.001)	0.23 (0.14 to 0.32; <0.001)	0.24 (0.15 to 0.32; <0.001)
	Group $\times$ Stage ⁺	-0.17 (-0.29 to -0.05; 0.007)	-0.21 (-0.34 to -0.08; 0.002)	-0.19 (-0.31 to -0.07; 0.003)

Statistical comparisons between groups and stages were performed using linear mixed-effects models adjusted for age and sex, with the resident group, history-only stage set as reference (n=172; 86 per group).

Data are estimated differences (Δ). Bootstrap resampling (2000 iterations) derived 95% confidence intervals and two-sided P values.

* Group: the difference between agent and resident at the history-only stage, with positive values indicating agent superior performance.

Stage: the change across assessment stages in the resident group, with positive values indicating resident beneficial.

⁺Group $\times$ Stage: the incremental gain across assessment stages between groups, with negative values indicating a smaller incremental gain for agent.

Supplementary Table 10. Comparison of Test Recommendation and Diagnostic Performance between high and low Medhistory quality cases

Outcome	Metric	Low quality (95% CI)	High quality (95% CI)	Difference (High – Low, 95% CI)	P value
Test recommendation	Recall	0.60 (0.56 to 0.65)	0.65 (0.59 to 0.70)	0.05 (-0.03 to 0.12)	0.20
	Precision	0.96 (0.94 to 0.99)	0.96 (0.94 to 0.99)	-0.001 (-0.04 to 0.04)	0.94
	F1	0.74 (0.70 to 0.78)	0.77 (0.74 to 0.81)	0.03 (-0.02 to 0.09)	0.23
Primary diagnosis	Recall	0.57 (0.47 to 0.66)	0.76 (0.69 to 0.84)	0.19 (0.07 to 0.32)	0.001
	Precision	0.88 (0.80 to 0.95)	0.92 (0.87 to 0.97)	0.04 (-0.05 to 0.13)	0.44
	F1	0.69 (0.60 to 0.77)	0.83 (0.78 to 0.89)	0.14 (0.04 to 0.25)	0.003

Data are micro-level (pooled) metrics across all cases in each quality group (n=172). Cases were stratified into high quality (MedHistory $\geq$ 70) and low quality (MedHistory<70). Bootstrap resampling (B = 2000) was used to derive 95% confidence intervals and two-sided P values. Positive differences (High – Low) indicate better performance for the high-quality group.

Supplementary Table 11. Comparison of Cost and Waiting Time between Groups

Outcome	Stage	Agent group	Agent group reference standard	Resident group	Resident group reference standard	Hodges-Lehmann Difference in Δ (Resident -Agent, 95% CI)	P value
Fee (CNY)	History-only	223.5 (90.0 to 554.0)	483.0 (134.0 to 598.0)	462.5 (174.0 to 638.0)	412.0 (159.0 to 623.0)	30.0 (3.0 to 70.0)	0.02
	Signs- integration	382.0 (134.0 to 598.0)		558.5 (174.0 to 638.0)		30.0 (21.0 to 87.0)	<0.001
Wait times (minutes)	History-only	56.2 (52.0 to 85.8)	85.8 (52.0 to 117.2)	85.8 (85.8 to 110.8)	85.8 (85.8 to 163.5)	0.0 (0.0 to 0.0)	0.92
	Signs- integration	85.8 (56.2 to 127.7)		85.8 (85.8 to 163.5)		0.0 (0.0 to 1.1)	<0.001

CNY = Chinese Yuan. Data for the agent, resident groups and their reference standard are median (IQR).

Group differences were assessed using the Mann–Whitney U test (n=172; 86 per group).

Hodges–Lehmann differences quantify between-group differences in absolute deviation from the reference standard (resident minus agent), with positive values indicating greater deviation in the resident group.

Supplementary Note 1 General Structure of Symptom-Oriented Logic Trees for Clinical History-Taking

Each logic tree is implemented as a directed acyclic graph that standardizes diagnostic inquiry. Key components include:

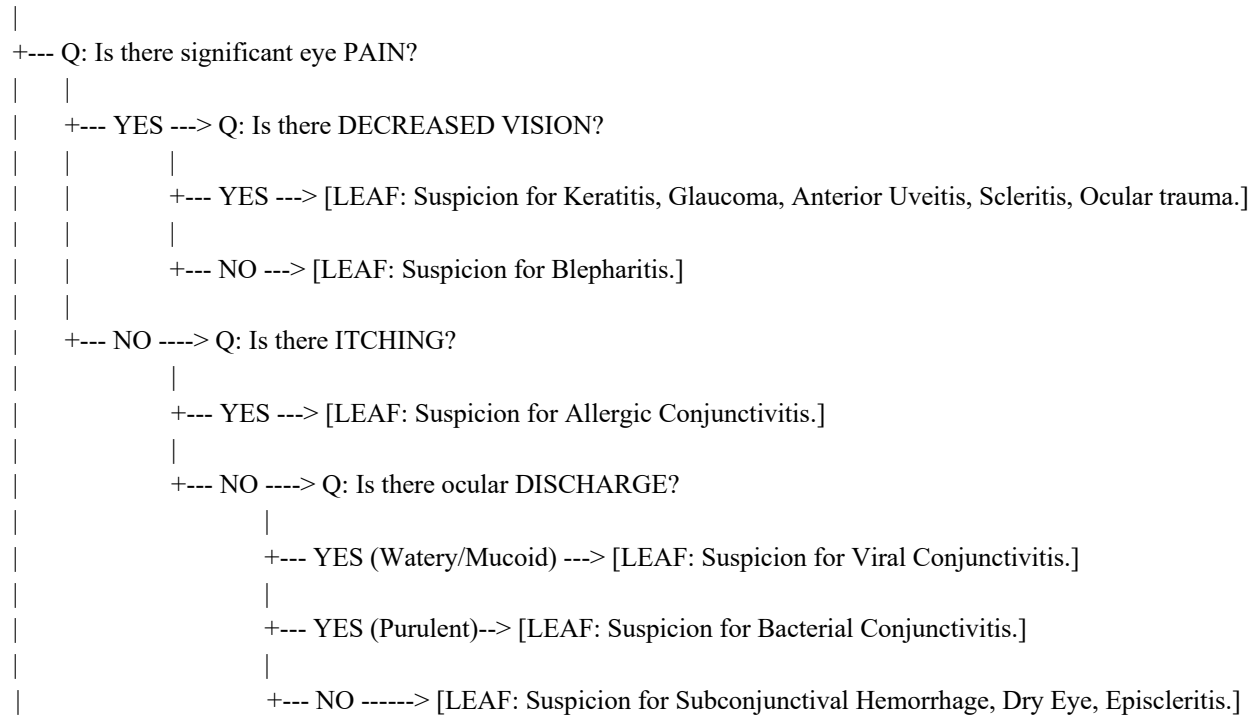
- **Root Node:** The patient's chief complaint (e.g., "Red Eye", "Vision Loss").
- **Decision Nodes:** Specific questions posed by the agent to elicit critical clinical features(e.g., "Was the vision loss sudden?" "Did any event precede these symptoms?"). Questions may be binary (Yes/No), multiple-choice, or open-ended.
- **Branches/Edges:** Pathways determined by the patient's response to a decision node.
- **Leaf Nodes:** Terminal points of an inquiry path, resulting in a differential diagnosis or a specific clinical recommendation.

Detailed Examples of Logic Trees

Text-based schematics for two of the four chief complaints are provided in the following sections.

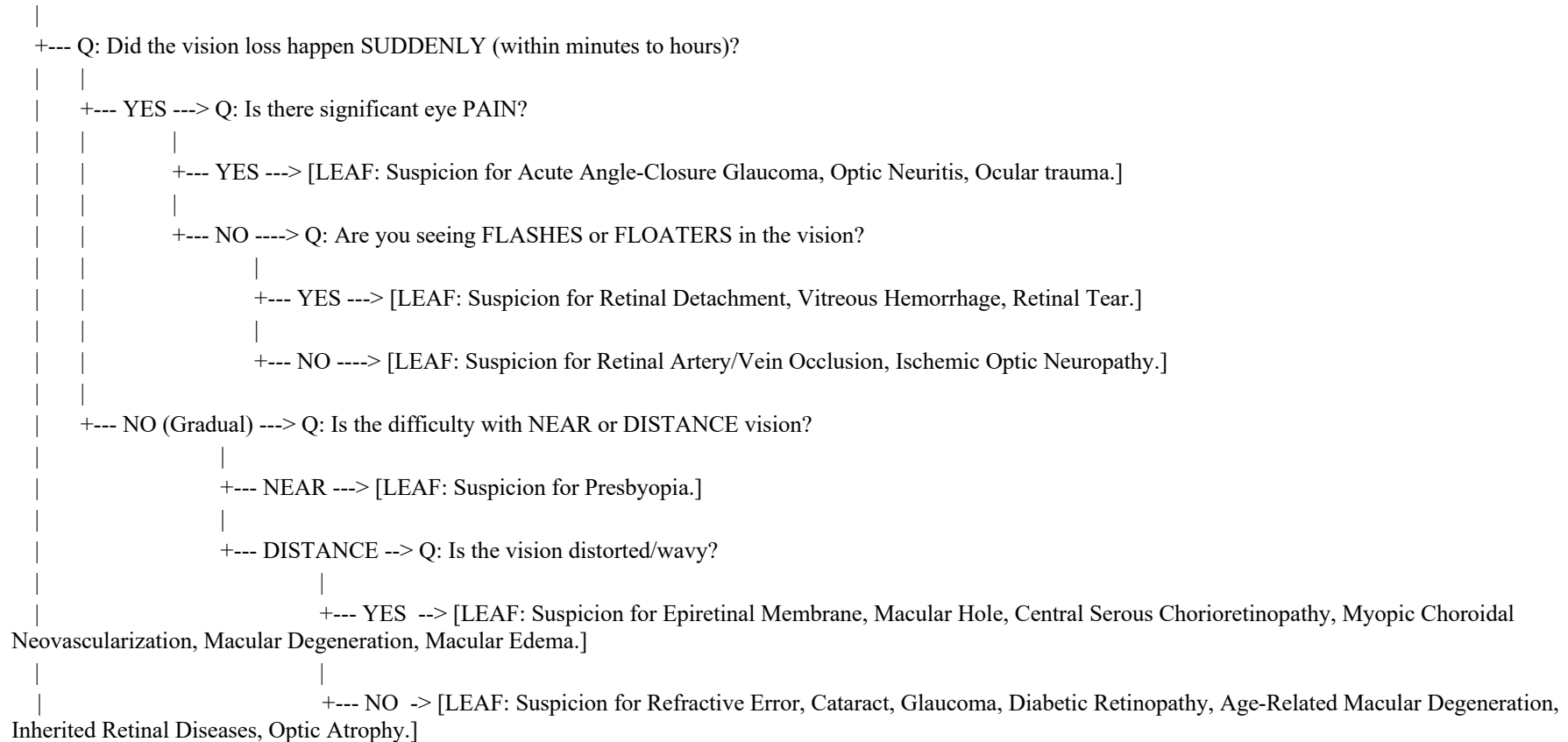
Example 1: Logic Tree for the Chief Complaint of “Red Eye”

[ROOT: Red Eye]



Example 2: Logic Tree for the Chief Complaint of “Vision Loss”

[ROOT: Vision Loss]



Summary of Logic Trees

The table below summarizes the scope and key characteristics of the four logic trees developed for this study.

Chief Complaint	Key Differentiating Questions	Examples of Potential Diseases
Red Eye	Pain, Vision Loss, Discharge, Itching, Dryness, etc.	Conjunctivitis, Uveitis, Glaucoma, Keratitis, Scleritis, Subconjunctival Hemorrhage, Dry Eye, etc.
Eye Pain	Localization (Superficial vs. Deep vs. Periocular), Vision Loss, Light Sensitivity, Tearing, Halos, etc.	Corneal Abrasion, Foreign Body, Uveitis, Glaucoma, Optic Neuritis, Scleritis, etc.
Vision Loss	Onset (Sudden/Gradual), Pain, Flashes/Floaters, Distortion, etc.	Retinal Detachment, Vascular Occlusion, Optic Neuritis, Cataract, Macular Degeneration, etc.
Floaters	Localization (Stationary vs. Mobile), Vision Loss, Flashes, Distortion, Color Perception, etc.	Vitreous Opacities, Posterior Vitreous Detachment, Retinal Tear/Detachment, Vitreous Hemorrhage, Optic Nerve and Visual Pathway Diseases, etc.

Note: The symptom-oriented logic trees are designed to guide history-taking rather than to serve as exhaustive frameworks.

Supplementary Note 2 Showcases

Examples for Model Comparison in the Proof-of-Concept Benchmarking

Patient profiles were developed from real-world outpatient cases by having residents role-play patients.

A 51-year-old female patient presents with a one-year history of blurred vision, photophobia, dry eye, and floaters in the left eye. Past ocular history includes dry eye, vitreous opacities and presbyopia. She underwent bilateral punctal plug insertion seven years ago for chronic dry eye. Systemic history includes cervical spondylosis.

All five models (Study Agent, Qwen-2.5 72B, GPT-4o, GPT-4.5, and DeepSeek-R1 671B) were evaluated using the same unified prompt described above.

Note: The original Chinese content has been translated into English and manually verified for accuracy.

Study Agent Pre-consultation

Study Agent: Hello, what brings you in today?

Patient: I experience blurred vision in my left eye, as if looking through fog. There is a central area appearing pixelated, and I am sensitive to light, which feels uncomfortable. Sometimes I need to squint to focus, and I often feel ocular dryness.

Study Agent: Understood. How long have these symptoms been present?

Patient: One year.

Study Agent: Was there any triggering factor when the symptoms first occurred?

Patient: No.

Study Agent: Have you ever had trauma to the left eye?

Patient: No.

Study Agent: Regarding the blurred vision, do you feel any vision loss?

Patient: I don't think my vision has changed much. Distance vision is still acceptable. Due to presbyopia, near vision can be blurry.

Study Agent: Have you noticed redness, pain, or new-onset floaters?

Patient: I have floaters.

Study Agent: Do the floaters remain fixed or move with eye movement?

Patient: They move with eye movement.

Study Agent: Have they increased in number? Any flashes, visual distortion, or color changes?

Patient: No.

Study Agent: Apart from ocular symptoms, do you experience systemic discomfort, such as headaches or nausea?

Patient: I have cervical spondylosis.

Study Agent: Have you previously sought medical care for this condition?

Patient: Yes.

Study Agent: Approximately how long ago was your last eye examination?

Patient: About one year ago.

Study Agent: What did the ophthalmologist suspect at that time? Any treatments given?

Patient: I attended a retina clinic and underwent several fundus examinations. No major issues were found except vitreous opacities. The doctor recommended observation and follow-up if symptoms worsened.

Study Agent: Have you used any eye drops? If yes, please specify the name.

Patient: No.

Study Agent: Any previous ocular diseases?

Patient: I have dry eye.

Study Agent: Any hypertension, diabetes, cardiovascular disease, or other systemic illnesses?

Patient: No.

Study Agent: Have you undergone any ocular surgeries? If so, which eye and what procedure?

Patient: I previously had dry eye and underwent bilateral punctal occlusion.

Study Agent: Approximately when was the procedure performed?
Patient: About seven years ago.
Study Agent: Do you wear glasses? Are they for distance, near, or both? What is the power of the lenses?
Patient: I wear reading glasses for near work.
Study Agent: Have you ever experienced allergic reactions to foods or medications, such as rash or shortness of breath?
Patient: No.
Study Agent: Do any of your immediate family members have similar eye diseases or hereditary disorders?
Patient: No.
Study Agent: Last question. Do you frequently use electronic devices? What kind of work do you do?
Patient: No, I am retired.

Study Agent: Thank you for your cooperation. Your medical record is as follows.

Chief Complaint:
 Blurred vision in the left eye with floaters for 1 year.

History of Present Illness:
 The patient developed insidious onset blurred vision in the left eye 1 year ago, accompanied by photophobia, ocular dryness, and mobile floaters, without eye pain, redness, or noticeable decline in visual acuity. No systemic symptoms were reported except for cervical spondylosis. One year ago, fundus examination at a retina clinic revealed vitreous opacities, and observation without active treatment was recommended. Symptoms have since progressively worsened.

Past Medical History:
 Ocular: Dry eye disease; vitreous opacities; wears presbyopic spectacles.
 Systemic: Cervical Spondylosis; denies other systemic illnesses.

Ocular Trauma History:
 Denies.

Surgical History:
 Bilateral punctal occlusion for dry eye approximately 7 years ago.

Allergy History:
 Denies known drug or food allergies.

Family History:
 Negative for hereditary or similar ocular conditions.

Personal History:
 Retired; does not use electronic devices frequently.

MedHistory Evaluation			
Section (Max Points)			Score and Rationale for Deduction (if any)
	Chief Complaint (30)		30
	History of Present Illness (30)	Precipitating factor (10)	10
		Characteristics of presenting complaint and associated symptoms (10)	10
		Clinical Course (5)	5
		Recent progression and evolution (5)	2.5; influencing factors not further explored.
	Past Medical	History of spectacle or contact lens wear (1)	0.5; duration/effectiveness not elicited

History-taking (92)	History (15)	General health status (4)	3; systemic medications not explored
		Trauma History (5)	5
		Ocular Surgical History (5)	5
	Family History (5)		5
	Allergy History (5)		5
	Personal History (5)		2.5; smoking/alcohol not addressed
	Logic Sequencing (2)		2
	Communication (4)		4
Documentation (4)		4	
Total (100)		93.5	

Qwen-2.5 72B Pre-consultation

Qwen: Hello, what brings you in today?

Patient: I experience blurred vision in my left eye, as if looking through fog. There is a central area appearing pixelated, and I am sensitive to light, which feels uncomfortable. Sometimes I need to squint to focus, and I often feel ocular dryness.

Qwen: How long have these symptoms been present?

Patient: One year.

Qwen: During this year, have the symptoms been persistent, or have they fluctuated?

Patient: They have worsened compared with before.

Qwen: Do you have any other associated symptoms, such as headache, eye pain, or vision loss?

Patient: I don't think my vision has changed much. Distance vision is still acceptable. Due to presbyopia, near vision can be blurry.

Qwen: Have you had any previous eye examinations or treatments?

Patient: I attended a retina clinic and underwent several fundus examinations. No major issues were found except vitreous opacities. The doctor recommended observation and follow-up if symptoms worsened.

Qwen: Do you use electronic devices for prolonged periods or perform tasks that require heavy near work?

Patient: Not really.

Qwen: Do you have a history of dry eye disease or use any medication that may cause ocular discomfort?

Patient: I had dry eye previously.

Qwen: Do you wear contact lenses?

Patient: No.

Qwen: Any history of allergies, such as to pollen, dust, or pet dander?

Patient: No.

Qwen: Any recent cold, fever, or systemic disease?

Patient: I have cervical spondylosis.

Qwen: Your medical record is as follows.

Patient information

Name: [Name]

Sex: [Sex]

Age: [Age]

Date: [Date]

Chief Complaint:

Blurred vision in the left eye with central pixelation, photophobia, and dryness for one year, recently worsened.

History of Present Illness:

The patient reports blurred vision in the left eye, with a central pixelated area, photophobia, and frequent dryness for one year, with recent symptom progression. Distance vision remains clear. Near vision is impaired due to presbyopia. One year ago, the patient attended the retina clinic and underwent retinal examinations that revealed vitreous opacities; observation was advised.

Past Medical History:

Dry eye disease. Cervical Spondylosis. No history of contact lens wear. No known allergies. Denies prolonged digital screen use. Denies other systemic illnesses.

Preliminary Impression:

Left-eye blurred vision with central visual disturbance, photophobia, dryness, and intermittent diplopia, possible progression of dry eye disease, vitreous opacities, or other posterior segment pathology.

Recommendations:

Comprehensive ophthalmic evaluation including IOP measurement, corneal topography, and optical coherence tomography (OCT). Continue artificial tear usage. Routine follow-up to monitor symptom evolution.

If you have any additional information or questions regarding the above medical record, please feel free to let me know.

MedHistory Evaluation

Section (Max Points)			Score and Rationale for Deduction (if any)
History-taking (92)	Chief Complaint (30)		30
	History of Present Illness (30)	Precipitating factor (10)	0; no known triggers asked
		Characteristics of presenting complaint and associated symptoms (10)	7.5; associated negative findings were not addressed.
		Clinical Course (5)	2.5; no specific timeline checkpoints
		Recent progression and evolution (5)	5
	Past Medical History (15)	History of spectacle or contact lens wear (1)	0.5; duration/effectiveness not elicited
		General health status (4)	2; systemic and ocular medications not explored
		Trauma History (5)	0; not asked
		Ocular Surgical History (5)	0; not asked
	Family History (5)		0; not asked
	Allergy History (5)		5
	Personal History (5)		2.5; smoking/alcohol and occupation not addressed
	Logic Sequencing (2)		2
Communication (4)			4
Documentation (4)			3; informal written style
Total (100)			64

GPT-4o Pre-consultation

GPT-4o: Hello, I am the ophthalmologist. What symptoms are troubling your eyes?

Patient: I experience blurred vision in my left eye, as if looking through fog. There is a central area appearing pixelated, and I am sensitive to light, which feels uncomfortable. Sometimes I need to squint to focus, and I often feel ocular dryness.

GPT-4o: Thank you for the detailed description. These symptoms sound somewhat complex, so I will ask questions step-by-step to understand more clearly. When did this begin?

Patient: One year ago.

GPT-4o: Understood. Have your symptoms changed over this year? Gradually worsening, fluctuating, or essentially stable?

Patient: They have worsened compared with before.

GPT-4o: Does your right eye have similar symptoms, or only the left eye is affected?

Patient: Only the left eye.

GPT-4o: Do you wear spectacles or contact lenses? Any myopia, hyperopia, or astigmatism?

Patient: I wear reading glasses for near work.

GPT-4o: Have you noticed eye redness, eye pain, increased discharge, or tenderness in the eyeball?

Patient: No.

GPT-4o: Do you have any systemic diseases such as hypertension, diabetes, or autoimmune disorders (e.g., Sjögren syndrome, lupus)?

Patient: I have cervical spondylosis.

GPT-4o: How are your visual habits? For example, prolonged screen time, extended near work, or irregular rest?

Patient: Fairly normal.

GPT-4o: Understood. Any prior ocular trauma, foreign body exposure, or strong-light injury? Any ocular surgeries?

Patient: I previously had dry eye and underwent a lacrimal duct probing procedure.

GPT-4o: Lastly, any family history of major ocular diseases such as glaucoma, macular degeneration, or retinal detachment?

Patient: No.

GPT-4o: Great, thanks for your cooperation. I will now document your medical record.

Chief Complaint:

Left-eye blurred vision with central pixelation for one year.

History of Present Illness:

The patient reports persistent blurred vision in the left eye for approximately one year, described as “looking through fog,” with a central pixelated area, photophobia, need to squint for focusing, and frequent dryness. No significant progression or relief. No symptoms in the right eye. No eye redness, pain, or discharge.

Past Medical History:

Dry eye disease; previous lacrimal duct probing. Cervical Spondylosis. No history of contact lens wear. No known allergies. No ocular trauma history. No other ocular surgeries.

Visual Habits: Fair overall, no prolonged excessive near work.

Family History: No significant ocular disease.

Preliminary Impression: Vitreous opacities. Persistent dry eye disease. Possible macular pathology requiring exclusion (e.g., epiretinal membrane, age-related macular degeneration). Refractive error and presbyopia may contribute.

Recommendations:

Macular optical coherence tomography (OCT). Anterior segment evaluation + tear breakup time (TBUT). Visual function testing (VA, contrast sensitivity, color vision). Bilateral IOP and vitreous evaluation. Advise regular visual rest, artificial tears as needed, and prompt follow-up if symptoms worsen or central vision declines.

If you'd like, I can continue refining the medical record — for example, by adding examination results or helping you format it into a standard electronic medical record (EMR). Is there anything else you think should be added?

MedHistory Evaluation

Section (Max Points)			Score and Rationale for Deduction (if any)	
History-taking (92)	Chief Complaint (30)		30	
	History of Present Illness (30)	Precipitating factor (10)	0; no known triggers asked	
		Characteristics of presenting complaint and associated symptoms (10)		10
		Clinical Course (5)		0; not asked
		Recent progression and evolution (5)		5
	Past Medical History (15)	History of spectacle or contact lens wear (1)		0.5; duration/effectiveness not elicited
		General health status (4)		2; systemic and ocular medications not explored
		Trauma History (5)		5
		Ocular Surgical History (5)		2.5; not asked date of surgery
	Family History (5)		0; not asked	
	Allergy History (5)		0; not asked	
	Personal History (5)		2.5; smoking/alcohol and occupation not addressed	
	Logic Sequencing (2)		2	
Communication (4)			4	
Documentation (4)			4	
Total (100)			62.5	

GPT-4.5 Pre-consultation

GPT-4.5: What symptoms are troubling your eyes?

Patient: I experience blurred vision in my left eye, as if looking through fog. There is a central area appearing pixelated, and I am sensitive to light, which feels uncomfortable. Sometimes I need to squint to focus, and I often feel ocular dryness.

GPT-4.5: How long have these symptoms been present? Did they start suddenly or gradually worsen?

Patient: One year. They have worsened compared with before.

GPT-4.5: Have you undergone any ophthalmic examinations or treatments within this year? If so, what were the results and were you given any diagnosis or recommendations?

Patient: I visited the retina clinic and had several fundus examinations. They did not find major abnormalities but noted some vitreous opacities. The doctor said to observe and return if symptoms worsen.

GPT-4.5: Do you have myopia, hyperopia, or astigmatism? Do you wear contact lenses?

Patient: I wear reading glasses for near work.

GPT-4.5: Aside from the ocular symptoms, do you have other systemic conditions such as headache, hypertension, diabetes, or other chronic diseases?

Patient: I have cervical spondylosis.

GPT-4.5: In your daily work or life, do you spend long hours on near tasks such as computer or smartphone use, or prolonged reading?

Patient: Not really.

GPT-4.5: The patient reports blurred vision as if viewed through fog, with a central mosaic-like disturbance, photophobia, and ocular dryness for one year. Prior examinations indicated vitreous opacities

without significant fundus abnormalities. The patient has a history of cervical spondylosis and presbyopia. Based on the available information, a comprehensive medical record is summarized below.

Chief Complaint:

Blurred vision in the left eye for one year.

History of Present Illness:

The patient reported a one-year history of decreased vision in the left eye, without clear triggering factors. Symptoms include fog-like blurring and a central mosaic-like scotoma, associated with photophobia that worsens discomfort under bright light. Squinting helps improve focus. Occasional diplopia and frequent dryness are noted. Previous retinal clinic evaluation showed vitreous floaters without definite fundus pathology; observation was recommended. Symptoms have remained stable without significant improvement or deterioration.

Past Medical History:

Cervical Spondylosis. No history of diabetes, hypertension, ocular trauma, or prior eye surgeries.

Visual Correction History: Presbyopia; uses reading glasses for near tasks. No contact lens wear.

Personal History: No prolonged near work. No smoking or alcohol exposure reported.

Preliminary Impression: Vitreous opacities left eye. Dry eye disease. Presbyopia. Possible macular pathology – further evaluation required (rule out macular degeneration/ERM).

Recommendations:

Slit-lamp examination; tear film stability and secretion testing. Macular OCT imaging; consider fundus fluorescein angiography if indicated. Artificial tears for symptom relief. Avoid prolonged visual strain; regular follow-up. Return sooner if visual deterioration occurs.

Physician Signature: [xxx]

Date: [YYYY/MM/DD]

The above constitutes the medical record compiled based on the information you provided. Please feel free to raise any questions or provide additional details if necessary.

MedHistory Evaluation

Section (Max Points)		Score and Rationale for Deduction (if any)
History-taking (92)	Chief Complaint (30)	30
	History of Present Illness (30)	Precipitating factor (10)
		0; no known triggers asked
		Characteristics of presenting complaint and associated symptoms (10)
		7.5; associated negative findings were not addressed.
	Past Medical History (15)	Clinical Course (5)
		5
		Recent progression and evolution (5)
		5
	Family History (5)	History of spectacle or contact lens wear (1)
		0.5; duration/effectiveness not elicited
		General health status (4)
		1; past eye history, systemic and ocular medications not explored
	Allergy History (5)	Trauma History (5)
		0; not asked
		Ocular Surgical History (5)
		0; not asked
	Personal History (5)	2.5; smoking/alcohol and occupation not addressed
	Logic Sequencing (2)	2

Communication (4)	4
Documentation (4)	2; precipitating factors, trauma, and surgery were not asked, only recorded as denied.
Total (100)	59.5

DeepSeek-R1 671B Pre-consultation	
DeepSeek-R1 671B: What discomfort are you experiencing in your eyes? Patient: I experience blurred vision in my left eye, as if looking through fog. There is a central area appearing pixelated, and I am sensitive to light, which feels uncomfortable. Sometimes I need to squint to focus, and I often feel ocular dryness. DeepSeek-R1 671B: When did these symptoms begin? Have you recently experienced any ocular trauma, worn contact lenses, or had exposure to chemical irritants? Patient: One year. I have not. DeepSeek-R1 671B: Besides the symptoms mentioned above, have you noticed any eye redness, pain, or increased discharge? Do you have diabetes, hypertension, or other systemic diseases? Any family history of similar ocular conditions? Patient: No. I have cervical spondylosis. DeepSeek-R1 671B: I would like to further understand your recent visual changes. Has the blurred vision in your left eye gradually worsened or remained relatively stable? Is it more noticeable during certain times of the day, such as morning or evening? Patient: They have worsened compared with before. DeepSeek-R1 671B: Based on your description, the progressive blurred vision in your left eye accompanied by central visual disturbance and photophobia raises concern for the following potential conditions: 1. Macular pathology Age-related macular degeneration (AMD): Patients may experience central scotoma or visual distortion (e.g., mosaic-like metamorphopsia) and photophobia. Early dry AMD can be asymptomatic, but visual acuity declines as the disease progresses. Epiretinal membrane / macular hole: Traction from a proliferative membrane on the retinal surface may lead to visual distortion and blurring. Central serous chorioretinopathy (CSC): More common among young to middle-aged adults, characterized by central vision reduction and metamorphopsia, often associated with stress. 2. Corneal disorders Keratoconus: Progressive corneal thinning and protrusion result in irregular astigmatism and monocular diplopia, leading to significant vision deterioration in advanced cases. Corneal dystrophy: Opacities in specific corneal layers impair corneal transparency and vision quality. 3. Other considerations Dry eye disease with refractive instability: Ocular surface dryness may exacerbate astigmatism or accommodative dysfunction, causing fluctuating visual clarity. Optic nerve or retinal vascular abnormalities: Such as ischemic optic neuropathy, although typically presenting with sudden vision loss. Recommendations: Urgent ophthalmic evaluation: comprehensive ophthalmic examination: including slit lamp biomicroscope, fundus examination, and intraocular pressure measurement. Optical coherence tomography (OCT): to assess macular structural integrity. Visual field testing / Amsler grid: to evaluate central field defects. Corneal topography: to rule out keratoconus or corneal shape abnormalities. Precautions and initial management: avoid eye rubbing; wear UV-protective sunglasses outdoors to reduce photophobia; use preservative-free artificial tears if dry eye is confirmed; reduce prolonged visual tasks and monitor whether symptoms correlate with visual fatigue. If these symptoms are caused by macular diseases, early intervention (e.g., anti-VEGF therapy or laser treatment) is crucial in preventing further deterioration. Delayed management may increase the risk of irreversible vision loss. Patient: generate a medical record.	

DeepSeek-R1 671B:**Name:** XXX **Sex:** X **Age:** XX **Date of Visit:** XXXX-XX-XX**Chief Complaint**

Progressive blurred vision in the left eye with central visual disturbance for 1 year, photophobia for 1 month.

History of Present Illness

The patient reports progressive blurred vision in the left eye for 1 year without obvious triggers and did not seek medical attention initially. Symptoms have worsened over the past 3 months. A “mosaic-like” scotoma and metamorphopsia (e.g., distortion of straight lines) developed in the central visual field. Photophobia emerged 1 month ago and is aggravated under strong light exposure. The patient denies eye redness, eye pain, increased discharge, headache, or nausea/vomiting. No history of ocular trauma or recent strenuous activity.

Past Medical History

Denies hypertension, diabetes, and autoimmune disorders. Denies long-term systemic medication use. Denies ocular surgery or drug allergy history.

Personal and Family History

Denies smoking and alcohol abuse. Denies family history of similar ocular diseases or inherited eye disorders.

Physical Examination (Simulated)

To be supplemented after clinical evaluation

Best-corrected visual acuity (BCVA): OD 1.0 | OS 0.5 (decreased vs. 1 year ago)

IOP: OD 16 mmHg | OS 18 mmHg

Slit lamp: Clear cornea, normal anterior chamber depth, round pupils, brisk light reflex

Direct ophthalmoscopy: Attenuated foveal reflex in the left eye, no obvious hemorrhage/exudation (OCT required)

Preliminary Impression (to be confirmed by imaging tests)

1. Macular pathology in the left eye

Age-related macular degeneration? Central serous chorioretinopathy? Epiretinal membrane?

2. Refractive error/keratoconus to be ruled out

3. Dry eye disease (requires tear function testing)

Recommendations:

1. Further ophthalmic examinations: Optical coherence tomography (OCT), Visual field testing (Humphrey or Amsler grid), Corneal topography, Fundus fluorescein angiography (if indicated).

2. Symptomatic treatment: Preservative-free artificial tears, Sunglasses for photoprotection

3. Follow-up within 3 days with completed tests; emergency visit for sudden vision deterioration.

Physician Signature: XXX

Note: This is a simulated record based solely on the reported symptoms. A definitive diagnosis requires a comprehensive ophthalmic evaluation including imaging.

MedHistory Evaluation

Section (Max Points)			Score and Rationale for Deduction (if any)
	Chief Complaint (30)		20, “photophobia for 1 month” was system-generated.
	History of Present Illness (30)	Precipitating factor (10)	0; no known triggers asked
		Characteristics of presenting complaint and associated symptoms (10)	10
		Clinical Course (5)	0; not asked
		Recent progression and evolution (5)	5
	History of spectacle or contact lens wear (1)		0; not asked

History-taking (92)	Past Medical History (15)	General health status (4)	1, past eye history, systemic and ocular medications not explored
		Trauma History (5)	5
		Ocular Surgical History (5)	0; not asked
	Family History (5)		5
	Allergy History (5)		0; not asked
	Personal History (5)		0; not asked
	Logic Sequencing (2)		1; lack of coherence of medical history questions
Communication (4)			1; multiple questions in one turn, required further prompting to generate medical documentation
Documentation (4)			1; trauma history, surgical history, and personal history were not inquired but were documented as “denied”; “Cervical spondylosis” was not documented.
Total (100)			49

Optimal Study Agent Pre-consultation in the Randomized Clinical Trial

Note: The original Chinese content has been translated into English and manually verified for accuracy.

Reference Standard Patient Key Facts Female, 55 years old. Left eye: Painless, progressive visual decline for six months. Right eye: History of retinal detachment, treated with retinal laser photocoagulation in 2024. Past medical history: myopia, -3.00 diopters; Arrhythmia (untreated); Oral progesterone medication for more than 2 years. Pertinent Negatives: Denies history of ocular trauma, family history and allergies. Required Tests Visual acuity, Intraocular pressure, Slit lamp biomicroscope, Fundus examination, Optical coherence tomography and Subjective refraction test for both eyes. Required Diagnoses Left eye cataract. Inappropriate Tests and Diagnoses Considered inappropriate according to expert opinion.
Study Agent Pre-consultation Study Agent: Hello, what brings you in today? Patient: I cannot see clearly with my left eye. Study Agent: I see, how long have you noticed blurred vision in your left eye? Patient: About six months. Study Agent: Was there any triggering factor when the symptoms first occurred? Patient: I work long hours and may overuse my eyes. Study Agent: Understood. Prolonged visual demand may indeed cause discomfort. Have you ever had any eye injuries? Patient: No. Study Agent: Based on your description, your left-eye vision has decreased. Have you noticed redness, pain, or new-onset floaters?

Patient: Sometimes I see a small dark spot.

Study Agent: Does this dot remain fixed or move with eye movement?

Patient: It moves with eye movement.

Study Agent: Have they increased in number? Any flashes, visual distortion, or color changes?

Patient: No.

Study Agent: Was the vision loss acute or gradual? Has it ever improved?

Patient: It worsened gradually.

Study Agent: Apart from ocular symptoms, do you experience systemic discomfort, such as headaches or nausea?

Patient: No.

Study Agent: Have you previously sought medical care for this condition?

Patient: Yes.

Study Agent: Approximately how long ago was your last eye examination?

Patient: About two months ago.

Study Agent: What did the ophthalmologist suspect at that time? Any treatments given?

Patient: They confirmed cataract in the left eye. I was given traditional Chinese medicine in another Hospital, but it was ineffective.

Study Agent: Have you used any eye drops? If so, please tell me the name.

Patient: Yes, but I forgot the name.

Study Agent: Apart from this cataract, have you had any other eye diseases?

Patient: Retinal detachment.

Study Agent: Which eye had the retinal detachment?

Patient: The right eye.

Study Agent: Do you have hypertension, diabetes, heart disease, or any other conditions?

Patient: I have arrhythmia.

Study Agent: Understood. Have you undergone any eye surgery?

Patient: Laser treatment in the right eye.

Study Agent: Could you please tell me which year this procedure was performed? If you don't recall the exact year, that's fine too.

Patient: Last year.

Study Agent: Do you wear glasses? Are they for distance, near, or both? What is the power of the lenses?

Patient: Myopic lenses.

Study Agent: Have you ever experienced allergic reactions to foods or medications, such as rash or shortness of breath?

Patient: No.

Study Agent: Do any of your immediate family members have similar eye diseases or hereditary disorders?

Patient: Nope.

Study Agent: Last question. Do you frequently use electronic devices? What kind of work do you do?

Patient: Yes, I am a teacher.

Study Agent: Your medical record is as follows.

Chief Complaint:

Decreased vision in the left eye for six months.

History of Present Illness:

The patient reported a six-month history of progressive visual decline in the left eye, which may relate to prolonged work hours and excessive near-work demand. The visual decline is accompanied by small, mobile, dot-like floaters that move with eye movements. The patient denies ocular pain or redness, denies systemic discomfort. Two months ago, the patient was evaluated elsewhere and diagnosed with left-eye cataract, for which traditional Chinese medicine was prescribed but with poor effect. The patient is currently using eye drops, although the specific medication is unknown. The patient presented to our clinic for further evaluation and management.

Past Medical History:

Ocular: wears myopic glasses.

Systemic: arrhythmia, denies other systemic illnesses.

Ocular Trauma History:

Denies.

Surgical History:

Right-eye laser photocoagulation performed last year for retinal detachment.

Allergy History:

Denies known drug or food allergies.

Family History:

Negative for hereditary or similar ocular conditions.

Personal History:

Teacher, use electronic devices frequently.

(Comment: oral medication history not obtained.)

Note: The following generated content is provided for research purposes only and will not be displayed in the user interface.

Study Agent: Initial Impression

Right eye: No obvious abnormalities noted.

Left eye: Suspected cataract, Suspected vitreous floaters.

Recommendations

Based on the patient's symptoms and medical history, left-eye cataract and vitreous floaters are suspected. The following examinations are recommended:

1. Slit-lamp biomicroscope – to assess lens opacity and possible vitreous abnormalities.
2. Fundus examination – to evaluate retinal status, particularly given the history of retinal detachment in the right eye.
3. Optical coherence tomography – to assess the macula and retina in greater detail.
4. Intraocular pressure measurement – to rule out glaucoma.

Additionally, regular monitoring of blood glucose and blood pressure is advised, as systemic diseases may affect ocular health. The patient is encouraged to maintain healthy visual habits, avoid prolonged near-work, and ensure adequate rest. If symptoms worsen or new symptoms appear, prompt ophthalmologic evaluation is recommended. Follow-up every 3–6 months is suggested.

Exploratory analysis (Ocular signs supplied in textual descriptions)

Examination Findings

Best-corrected visual acuity: OD 1.0, OS 0.15

Non-contact intraocular pressure: OD 14.7 mmHg, OS 14 mmHg

Examination	Right Eye	Left Eye
Eyelids	No swelling or mass	No swelling or mass
Conjunctiva	No hyperemia or edema	No hyperemia or edema
Cornea	Clear, no KP	Clear, no KP
Anterior chamber	Normal depth, clear	Normal depth, clear
Iris & pupil	Clear iris pattern, round pupil	Clear iris pattern, round pupil
Lens	Clear	Posterior subcapsular opacity
Fundus	Retina flat under small pupil	Poor view under small pupil
Others	–	–

Study Agent:

{ "Reason": "Based on the pre-consultation report and examination findings, the patient's left eye shows significant visual decline with floaters and mild photophobia. The fundus view is obscured under a small pupil, and posterior subcapsular cataract is present, strongly suggesting cataract as the primary cause.

Given the history of right-eye retinal detachment, recurrent or related pathology must also be considered. Although right-eye vision is currently normal, further evaluation is warranted.”

“OD_Report”: “1. No obvious abnormalities.”,

“OS_Report”: “1. Suspected cataract. 2. Suspected vitreous floaters. 3. Retinal pathology cannot be ruled out.”,

“examination”: “1. Subjective refraction test OU; 2. B-scan ultrasonography OS; 3. OCT OU; 4. Fundus examination OU.”,

“Treatment”: “1. Cataract surgery is recommended for the left eye. 2. Continue monitoring the right eye and seek prompt evaluation if symptoms occur.”}

Suboptimal Study Agent Pre-consultation in the Randomized Clinical Trial

Note: The bellowing content originally in Chinese were translated into English and manually reviewed for accuracy. The text marked in red is the comment from the evaluators.

Reference Standard

Patient Key Facts

Female, 62 years old.

Left eye: Postoperative foreign-body sensation and dryness for 2 weeks following cataract surgery.

Right eye: Cataract surgery performed 1 month ago.

Past medical history: presbyopia, +2.00 diopters.

Pertinent Negatives: Denies history of systemic disease, ocular trauma, family history and allergies.

Required Tests

Visual acuity, Intraocular pressure, Slit lamp biomicroscope, Fundus examination, Ocular surface staining and Subjective refraction test for both eyes.

Required Diagnoses

Left eye: dry eye.

Inappropriate Tests and Diagnoses

Considered inappropriate according to expert opinion.

Study Agent Pre-consultation

Study Agent: Hello, what brings you in today?

Patient: The operated eye often has a foreign-body sensation, and the wound area occasionally feels achy or pressured.

Study Agent: Could you please tell me which eye underwent surgery?

Patient: Left eye.

Study Agent: How long has your left eye been experiencing these symptoms?

Patient: The foreign-body sensation has been present since the surgery.

Study Agent: Understood. Has your left eye ever been injured?

Patient: No.

Study Agent: After the surgery, have you noticed any decrease in vision in your left eye?

Patient: I don't feel vision loss in daily life, but the one-week postoperative check showed worse visual acuity compared with the day after surgery. And I feel slight blurriness when looking at things within about 2 meters.

Study Agent: Have you noticed redness, pain, or new-onset floaters?

Patient: Sometimes it looks a bit red.

Study Agent: Have you noticed increased discharge or dryness in your eyes recently?

Patient: My eyes do feel dry.

Study Agent: Apart from ocular symptoms, do you experience systemic discomfort, such as headaches or nausea?

Patient: No.

Study Agent: From your description, you have postoperative discomfort. Have you sought any medical care for this condition before today?

Patient: Not yet.

Study Agent: Have you used any eye drops? If so, please tell me the name. Patient: The eye drops were all prescribed by the doctor upon discharge. Study Agent: Aside from this episode, have you ever had other eye diseases? Patient: I don't think so. Study Agent: Do you have hypertension, diabetes, heart disease, or any other conditions? Patient: No. Study Agent: Do you wear glasses? Are they for distance, near, or both? What is the power of the lenses? Patient: I only use reading glasses for near work, about +2.00 diopters. Study Agent: Have you ever experienced allergic reactions to foods or medications, such as rash or shortness of breath? Patient: No. Study Agent: Do any of your immediate family members have similar eye diseases or hereditary disorders? Patient: Nope. Study Agent: Last question. Do you frequently use electronic devices? What kind of work do you do? Patient: No, I am retired. Study Agent: Your medical record is as follows. Chief Complaint: Foreign-body sensation, redness and dryness in the left eye for over 10 days. (Comment: "10 days" is speculated; the surgery details such as the exact date and specific procedure are not inquired.) History of Present Illness: The patient has experienced persistent foreign-body sensation in the left eye following surgery, accompanied by redness, dryness, and near-vision blur for over 10 days. The patient denies any decrease in visual acuity or increased ocular discharge. The patient has been regularly using postoperative eye drops and presented to our hospital for further treatment. Past Medical History: Ocular: wears presbyopic spectacles of approximately +2.00 D for near tasks. Systemic: denies systemic illnesses. Ocular Trauma History: Denies. Surgical History: History of left-eye surgery; the exact surgical procedure is unknown. Allergy History: Denies known drug or food allergies. Family History: Negative for hereditary or similar ocular conditions. Personal History: Retired, denying use electronic devices frequently. <i>Note: The following generated content is provided for research purposes only and will not be displayed in the user interface.</i> Study Agent: Initial Impression Right eye: No obvious abnormalities noted. Left eye: Suspected dry eye disease, suspected postoperative complication to be further evaluated. Recommendations Based on the patient's symptoms, dry eye disease is suspected, although postoperative complications cannot be excluded. A detailed ophthalmic evaluation, particularly 1.slit-lamp examination and 2.tear function tests is recommended to establish a definitive diagnosis. If dry eye disease is confirmed, artificial tears may be prescribed, and environmental/lifestyle modifications advised. If postoperative complications are identified, further management will depend on the clinical findings. The patient is encouraged to maintain proper visual hygiene and revisit every 3–6 months. Exploratory analysis (Ocular signs supplied in textual descriptions) Examination Findings

Unaided visual acuity: OD 0.8, OS 0.63

Non-contact intraocular pressure: OD 14.7 mmHg, OS 14.7 mmHg

Examination	Right Eye	Left Eye
Eyelids	No swelling or mass	No swelling or mass
Conjunctiva	No hyperemia or edema	No hyperemia or edema
Cornea	Clear, no KP	Clear, no KP
Anterior chamber	Normal depth, clear	Normal depth, clear
Iris & pupil	Clear iris pattern, round pupil	Clear iris pattern, round pupil
Lens	Intraocular lens in place	Intraocular lens in place
Fundus	No abnormalities under small pupil	No abnormalities under small pupil
Others	—	—

Study Agent:

{"Reason": "Based on the pre-consultation report and clinical findings, the patient primarily presents with foreign-body sensation, ocular soreness, redness, dryness, and blurred near vision in the left eye. Intraocular pressure is within normal range, and non-dilated slit-lamp examination shows no significant abnormalities. Given the patient's age and symptom profile, dry eye disease is the most likely diagnosis, while mild conjunctivitis or subtle corneal surface abnormalities should also be considered. The right eye shows no symptoms or abnormal findings at present.",

"OD_Report": "1. No significant abnormalities detected.",

"OS_Report": "1. Suspected dry eye disease. 2. Suspected mild conjunctivitis. 3. Suspected corneal surface abnormality.",

"examination": "1. Schirmer test OU. 2. Tear film break-up time OU. 3. Fundus examination OU.",

"Treatment": "1. Use artificial tears to relieve symptoms of dry eye. 2. If conjunctivitis is confirmed, apply topical antibiotic or antiviral eye drops as appropriate. 3. Maintain good ocular hygiene and avoid prolonged visual tasks. 4. Schedule regular follow-up to monitor clinical changes."

}

(Comment: The physical signs were not fully understood. The finding of no conjunctival congestion or edema, and a clear cornea with no keratic precipitates, should rule out conjunctivitis and corneal abnormalities.)

Optimal Resident Pre-consultation in the Randomized Clinical Trial

Note: The bellowing content originally in Chinese were translated into English and manually reviewed for accuracy. The text marked in red is the comment from the evaluators.

Reference Standard

Patient Key Facts

Female, 61 years old.

Left eye: Painless, progressive visual decline for over six months; History of primary angle-closure glaucoma, central retinal vein occlusion.

Right eye: Cataract; History of primary angle-closure glaucoma.

Surgical history: Both eyes underwent laser peripheral iridotomy; Left eye underwent retinal laser photocoagulation, followed by cataract surgery one year ago.

Past medical history: Type 2 Diabetes Mellitus, controlled by medication.

Pertinent Negatives: Denies history of ocular trauma, family history and allergies.

Required Tests

Visual acuity, Goldmann applanation tonometry, Slit lamp biomicroscope, Fundus examination, Optical coherence tomography, Visual field testing and Subjective refraction test for both eyes.

Required Diagnoses

Bilateral primary angle-closure glaucoma, Left eye central retinal vein occlusion, Right eye cataract. Inappropriate Tests and Diagnoses Considered inappropriate according to expert opinion.
Resident Pre-consultation Resident: What eye symptoms are you experiencing? Patient: I feel that my vision has decreased. Resident: Which eye is affected? Patient: The left eye. Resident: How long has this been going on? Patient: More than six months. Resident: Any other discomfort, such as redness or pain? Patient: No. Resident: Have you had any previous eye diseases? Patient: I have glaucoma, and retinal vessel blockage. Resident: Which eye was affected? Patient: Both eyes have glaucoma, and the left eye had retinal vein occlusion. Resident: Have you received any treatments? Patient: Both eyes underwent laser iridotomy, and only the left eye had retinal laser one year ago. Resident: Are you using any eye drops now? Patient: For the left eye, I use brimonidine, timolol, and tafluprost. Resident: Do you have any systemic diseases such as hypertension or diabetes? Patient: I have diabetes. Resident: Besides what you mentioned, have you had any other eye surgeries? Patient: I had cataract surgery on the left eye last year. Resident: Any drug allergies? Patient: None. Medical record is as follows. Chief Complaint: Progressive visual decline in the left eye for more than six months. History of Present Illness: The patient's visual acuity gradually decline over six months, denies eye redness or pain and presented to our hospital for further evaluation. Past Medical History: Ocular: left eye retinal vein occlusion, bilateral glaucoma (brimonidine, timolol, and tafluprost for left eye). Systemic: diabetes mellitus, denies other systemic illnesses. Surgical History: Left eye cataract surgery and retinal laser photocoagulation were performed last year. Laser peripheral iridotomy in both eyes for glaucoma. Allergy History: Denies drug allergies. (Comment: ocular trauma history is missed.)
Resident: Initial Impression Bilateral glaucoma, left eye retinal vein occlusion, diabetes mellitus. Recommendations Visual acuity, Intraocular pressure, Slit lamp biomicroscope, Fundus examination, Subjective refraction test, Visual field testing, Optical coherence tomography for both eyes.
Exploratory analysis (Ocular signs supplied in textual descriptions) Examination Findings Unaided visual acuity: OD 0.63, OS 0.025 Non-contact intraocular pressure: OD 15.3 mmHg, OS 19.3 mmHg

Examination	Right Eye	Left Eye
Eyelids	No swelling or mass	No swelling or mass
Conjunctiva	No hyperemia or edema	No hyperemia or edema
Cornea	Clear, no KP	Clear, no KP
Anterior chamber	Central 3 CT, peripheral 1/4 CT, clear aqueous	Central 3 CT, peripheral 1/4 CT, clear aqueous
Iris & pupil	Clear iris pattern, Patent peripheral iridotomy, round pupil	Clear iris pattern, Patent peripheral iridotomy, round pupil
Lens	Opacity	Intraocular lens in place
Fundus	Vitreous opacity; optic disc with normal color and clear margins, C/D=0.3; retina attached; macula unremarkable.	Vitreous opacity; pale optic disc, C/D=1.0; retina attached; macula unremarkable; flame-shaped retinal hemorrhages; laser scars are visible.
Others	—	—
Resident: Primary diagnosis: Bilateral glaucoma, left eye central retinal vein occlusion, right eye cataract, diabetes mellitus. Examination: Binocular subjective refraction test, Visual field testing, Optical coherence tomography. Treatment: Control intraocular pressure; conservative management.		

Suboptimal Resident Pre-consultation in the Randomized Clinical Trial

Note: The following content originally in Chinese were translated into English and manually reviewed for accuracy. The text marked in red is the comment from the evaluators.

Reference Standard Patient Key Facts Male, 35 years old. Bilateral visual decline with a fixed scotoma for 8 days. No history of previous medical visits. Past medical history: myopia, -3.00 diopters; Type 2 Diabetes Mellitus, controlled by insulin. Pertinent Negatives: Denies history of surgical history, ocular trauma, family history and allergies.
Required Tests Visual acuity, Intraocular pressure, Slit lamp biomicroscope, Fundus examination, B-scan ultrasonography, Subjective refraction test for both eyes.
Required Diagnoses Bilateral vitreous hemorrhage, proliferative diabetic retinopathy.
Inappropriate Tests and Diagnoses Considered inappropriate according to expert opinion.
Resident Pre-consultation Resident: What eye symptoms are you experiencing? Patient: I feel that my vision has decreased. Resident: Which eye is affected? Patient: Both eyes. Resident: How long has this been going on? Patient: Around eight days. Resident: Any other discomfort, such as redness or pain? Patient: It feels like something is floating in front of my eye. Resident: Have you had any previous eye diseases?

Patient: Nope.

Resident: Do you have any systemic diseases such as hypertension or diabetes?

Patient: I have diabetes.

Medical record is as follows.

Chief Complaint:

Decreased vision in both eyes accompanied by floaters for 8 days.

History of Present Illness:

The patient reports decreased vision in both eyes accompanied by floaters for 8 days and presents to our hospital for evaluation.

Past Medical History:

Ocular: Denies.

Systemic: diabetes mellitus, denies other systemic illnesses.

(Comment: surgical history, ocular trauma history, allergy history is missed.)

Resident: Initial Impression

Bilateral diabetic retinopathy.

Recommendations

Visual acuity, Intraocular pressure, Slit lamp biomicroscope, Fundus examination, Subjective refraction test, Optical coherence tomography for both eyes.

Exploratory analysis (Ocular signs supplied in textual descriptions)

Examination Findings

Unaided visual acuity: OD 0.05, OS 0.01

Non-contact intraocular pressure: OD 16 mmHg, OS 15 mmHg

Examination	Right Eye	Left Eye
Eyelids	No swelling or mass	No swelling or mass
Conjunctiva	No hyperemia or edema	No hyperemia or edema
Cornea	Clear, no KP	Clear, no KP
Anterior chamber	Central 3 CT, peripheral 1/4 CT, clear aqueous	Central 3 CT, peripheral 1/4 CT, clear aqueous
Iris & pupil	Clear iris pattern, Patent peripheral iridotomy, round pupil	Clear iris pattern, Patent peripheral iridotomy, round pupil
Lens	Clear	Clear
Fundus	Vitreous hemorrhagic opacity; optic disc with normal color and clear margins, C/D = 0.4; scattered retinal hemorrhages; macula unremarkable.	Vitreous hemorrhagic opacity; optic disc with normal color and clear margins, C/D = 0.4; fibroproliferative membrane noted in the posterior pole; macula unremarkable.
Others	—	—

Resident:

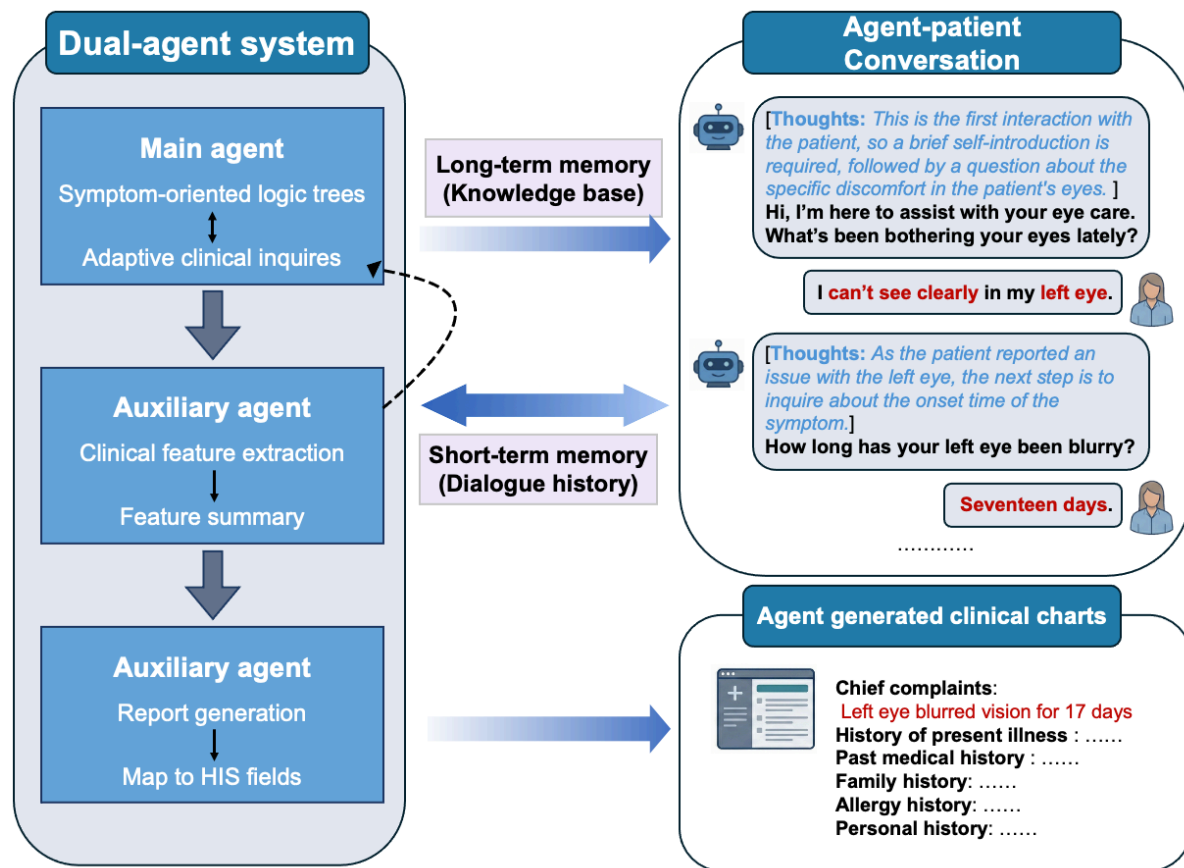
Primary diagnosis: Bilateral proliferative diabetic retinopathy.

Examination: Visual acuity, Intraocular pressure, Slit lamp biomicroscope, Fundus examination, Subjective refraction test, Optical coherence tomography, B-scan ultrasonography for both eyes.

Supplementary Note 3 Speech-to-text (STT) artifacts

For STT engine, patient speech was transcribed using the iFLYTEK Voice Dictation streaming WebAPI (<wss://iat-api.xfyun.cn/v2/iat>). When voice output was required, LLM-generated responses were delivered as synthesized speech using the iFLYTEK Text-to-Speech (TTS) streaming WebAPI (<wss://tts-api.xfyun.cn/v2/tts>). Model ID is M_CC164FF7040141CD88C96B59B90EC695.

We performed a post hoc manual review of a random sample of 50 participants. Across 24,923 characters in the speech-to-text (STT) records, 922 character errors were identified, corresponding to 3.7 artifacts per 100 characters. In the final clinical charts used for MedHistory scoring, 74 character errors were identified across 14,760 characters, corresponding to 0.5 artifacts per 100 Chinese characters.



Supplementary Figure 1. Architecture of the Dual-Agent System with Symptom-Oriented Reasoning

The system comprises two coordinated agents: a **Main Agent** conducts adaptive, multi-turn dialogues with the patient, guided by symptom-oriented logic trees to perform context-aware clinical inquiry; an **Auxiliary Agent** extracts and summarizes key clinical features from the ongoing interaction, updates the conversational context, and generates a semi-structured clinical note compatible with the hospital information system (HIS). A memory module supports both agents by storing domain knowledge, dialogue history, and patient-specific features for real-time retrieval. The entire interaction is orchestrated using chain-of-thought reasoning, enabling stepwise clinical reasoning throughout the pre-consultation history-taking.

Supplementary Note 4 Large Language Model Specifications and Inference

Parameters

We benchmarked the agent against state-of-the-art LLMs between April 22 and 30, 2025, with model memory and chat history functions disabled. Each dialogue turn was processed independently to restrict model responses to the current context only. Models parameters were reported following the TRIPOD-LLM checklist¹.

(1). Study Agent

- **Model Identification:**
 - **Model Name:** Study Agent
 - **Base Model:** Qwen-2.5-72B Instruct
 - **Customization:** Fine-tuned using ophthalmology records as previously published^{2,3}.
 - **Developer:** Zhongshan Ophthalmic Center
- **Inference Parameters:**
 - **Access Method:** Local deployment and inference
 - **Temperature:** 0.3
 - **Top-p:** 0.8
 - **Maximum Output Tokens:** 4096
 - **Frequency / Presence Penalty:** 1.05

(2). GPT-4o & GPT-4.5

- **Model Identification:**
 - **Model Name:** gpt-4o-2024-05-13 and gpt-4.5-preview-2025-02-27
 - **Developer:** OpenAI
 - **Model website:** <https://chatgpt.com/>
- **Inference Parameters:**
 - **Access Method:** OpenAI API
 - **Temperature:** 0.6
 - **Top-p:** 0.9
 - **Maximum Output Tokens:** 4096
 - **Frequency / Presence Penalty:** 0.0

(3). Qwen-2.5 72B

- **Model Identification:**
 - **Model Name:** Qwen-2.5 72B Instruct
 - **Developer:** Alibaba Cloud
 - **Model website:** <https://huggingface.co/Qwen/Qwen2.5-72B-Instruct>

- **Inference Parameters:**
 - **Access Method:** Local deployment and inference
 - **Temperature:** 0.6
 - **Top-p:** 0.9
 - **Maximum Output Tokens:** 4096
 - **Frequency / Presence Penalty:** 0.0

(4). DeepSeek-R1 671B

- **Model Identification:**
 - **Model Name:** DeepSeek-R1 671B
 - **Developer:** DeepSeek AI
 - **Model website:** <https://huggingface.co/deepseek-ai/DeepSeek-R1>
- **Inference Parameters:**
 - **Access Method:** Local deployment and inference
 - **Temperature:** 0.6
 - **Top-p:** 0.9
 - **Maximum Output Tokens:** 4096
 - **Frequency / Presence Penalty:** 0.0

Supplementary Table 12. MedHistory Framework for Ophthalmic History-taking and Documentation

Section (Points)		Content		Scoring Criteria
History-taking (92)	Chief Complaint (30)	Main symptoms, duration, laterality		30 points total, 0 if missing. • Symptoms: 10 points for clear and specific description (e.g. “sudden onset of blurred vision”); partial credit if vague (e.g., “blurred vision”). • Duration: 10 points for a precise timeline (e.g. “3 days”); partial credit if only approximate (e.g., “recent,” “for a while”) . • Laterality: 10 points if side (left, right, or bilateral) is clearly specified; partial credit if vague (e.g. “one eye”) .
	History of Present Illness (30)	Precipitating factor		10 points: identifies likely causes, triggers or risk factors; if none, states “no known precipitating factors”.
		Characteristics of presenting complaint and associated symptoms (positive and negative findings)		10 points total. 5 points if characteristics of presenting complaint are addressed (e.g. “acute or chronic”); 5 points for eliciting both positive and relevant negative symptoms (symptoms absent which are useful for differential diagnosis). 2.5 points if only positive or only negative symptoms are addressed.
		Clinical Course	Time	1 point: dates/timelines for key events.
			Diagnosis	1 point: mention any provisional or prior diagnoses.
			Treatment and response	3 points: any treatments tried and whether they helped (improved / unchanged / worsened).
		Recent progression and evolution	Severity (worsening/alleviation), frequency (increase/decrease) and influencing factors	5 points total. 2.5 points if partially mentioned.
	Past Medical History (15)	History of spectacle or contact lens wear	type of correction, duration and effectiveness	1 point total. 0.5 points if partially mentioned.
		General health status	Systemic disease, past ocular disease, and current medications (name and duration)	4 points total. 1 point for each content.
		Trauma History	Injury site and mechanism, diagnostic and therapeutic process	5 points total. Partial mention merits 2.5 points.
		Ocular Surgical History	Surgery date, laterality, procedure	5 points total. Partial mention merits 2.5 points.

	Family History (5)	Hereditary diseases, family history of similar conditions, relationship to the patient	5 points total. Partial mention merits 2.5 points.
	Allergy History (5)	Food/drug allergies, allergy symptoms	5 points total. Partial mention merits 2.5 points.
	Personal History (5)	Eye-use habits, smoking/alcohol exposure, occupation	5 points total. Partial mention merits 2.5 points.
	Logic Sequencing (2)	Logical order and coherence of medical history questions	2 points if medical history is obtained in a logical and coherent sequence.
Communication (4)	Empathy		2 points: use patient-centered language; show respect and acknowledges patients' feelings.
	Clarity		2 points: use understandable language, no repetition requested by the patient.
Documentation (4)	Correct use of medical terminology		2 points: convert patients' spoken language into formal medical terminology.
	Consistency of dialogue		2 points: documentation matches what patient said, no contradictions, history is coherent in record.

Supplementary Note 5 Prompt Incorporated Ocular Signs in the Randomized Clinical Trial

Textual descriptions of ocular signs (slit-lamp findings, visual acuity, and intraocular pressure) were provided to both groups after the pre-specified outcomes in the History-only stage were assessed. The following prompt was used for Study Agent.

Note: The prompts originally in Chinese were translated to English for demonstration purpose.

You are an ophthalmologist, skilled in patient interviewing and medical record documentation. Your task is to infer the patient's possible ocular diseases based on the following information: sex, age, visual acuity, intraocular pressure, pre-consultation chart, and slit-lamp findings. Generate the final report and recommendations following the Report Template below.

##Sex##

{xx}

##Age##

{xx}

##Visual acuity##

Right eye: {}

Left eye: {}

##Intraocular pressure##

Right eye: {}

Left eye: {}

##Pre-consultation chart##

{xxxxx}

##Slit-lamp findings##

Examination Site	Right Eye	Left Eye
-----	-----	-----
Eyelid		
Conjunctiva		
Cornea		
Anterior chamber		
Iris and pupil		
Lens		
Fundus		

##Reporting Reference Template##

No significant abnormality detected.

Suspected refractive error.

Suspected dry eye disease.

Suspected pterygium.

Suspected subconjunctival hemorrhage.

Other ocular conditions to be further evaluated.

##Report Template##

1. Output must strictly follow JSON format. No keys other than those listed in the JSON structure under “Report Template” are allowed.

2. Markdown formatting is not permitted.

3. JSON structure { {

 "Reason": "Provide a detailed explanation of the reasoning process used to infer potential diagnoses.",

 "OD_Report": "Based on the pre-consultation report, ocular signs, and visual acuity/IOP findings, list all possible conditions for the right eye. Follow the format and style in the Reporting Reference Template.

 Number each condition starting from 1. If no abnormalities are suspected, output: 1. No significant abnormality detected.",

 "OS_Report": "Based on the pre-consultation report, ocular signs, and visual acuity/IOP findings, list all possible conditions for the left eye. Follow the format and style in the Reporting Reference Template.

 Number each condition starting from 1. If no abnormalities are suspected, output: 1. No significant abnormality detected.",

 "examination": "Based on the inferred conditions, provide concise examination recommendations in a clinician's tone, e.g., 1. Cycloplegic refraction; 2. Anterior segment photography.",

 "Treatment": "Based on the inferred conditions, provide concise treatment recommendations in a clinician's tone, e.g., 1. Spectacle prescription; 2. Conservative medical therapy." } }

Supplementary Table 13. Qualitative Evaluation Metrics

Stage	Test recommendation rating scale	Primary diagnosis rating scale	Treatment recommendation rating scale
History-only	0=Completely inconsistent with history or pose a clear safety risk. 1=Key omission of 1–2 necessary tests to confirm/rule out the main diagnosis or an emergency or suggests tests without adequate historical basis. 2=Mostly reasonable but omits 1–2 relevant tests or partially deviates from the correct diagnostic pathway. 3=Recommendations are fully appropriate and logically support the complete diagnostic process.	0=Completely incorrect or critically misleading diagnosis (e.g., ignoring red flag symptoms or no plausible connection to the chief complaint). 1=Key omission of a core diagnostic clue evident in the history, but the general direction is plausible. 2=Mostly correct and grounded in the history but includes a significant omission or is suboptimal. 3=Optimal diagnostic accuracy achievable based purely on the patient history provided.	Not applicable
Signs-integration	0= Recommendations completely inconsistent with history and signs or pose a clear safety risk. 1=Key omission of 1–2 necessary tests to confirm/rule out the main diagnosis or an emergency, or suggests tests based on irrelevant information. 2=Mostly reasonable but omits 1–2 relevant tests or is partially insufficient to support the complete diagnostic process. 3=Recommendations are fully appropriate and logically support the complete diagnostic process.	0=Diagnosis fundamentally contradicts signs or history or leads to a critical safety error. 1=Key omission of a diagnosis directly implied by ≥ 1 clinical sign/symptom, or the diagnosis is not sufficiently supported by all available data. 2=Mostly correct but includes omissions or imprecision (e.g., failing to filter out less likely differentials) or misses 1–2 key points. 3=Optimal diagnostic accuracy achievable when integrating both history and clinical signs.	0=Treatment is critically incorrect or majorly contradicts the primary diagnosis (e.g., proposing surgery for a benign self-limiting condition). 1=Treatment is functional but inadequate to resolve the patient's issue or fails to address a clear surgical/emergency requirement. 2=The treatment direction is logical, but the judgment or the completeness of the strategy is marginally lacking. 3=Appropriately handles the treatment principles and protocols (e.g., urgency, conservative/deferred management, or appropriate referral strategy).

Supplementary References

- 1 Gallifant J, Afshar M, Ameen S, *et al.* The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med* 2025; **31**: 60–9.
- 2 Luo M-J, Pang J, Bi S, *et al.* Development and Evaluation of a Retrieval-Augmented Large Language Model Framework for Ophthalmology. *JAMA Ophthalmol* 2024; **142**: 798.

3 Luo M-J, Bi S, Pang J, *et al.* A large language model digital patient system enhances ophthalmology history taking skills. *Npj Digit Med* 2025; **8**: 502.

The CONSORT Reporting Checklist

	Item Description	Location (or reason for not reporting)
Title and Abstract		
1a. Title	Identification as a randomised trial.	Page 1
1b. Structured Abstract	Structured summary of the trial design, methods, results, and conclusions.	Page 2
Open Science		
2. Trial Registration	Name of trial registry, identifying number (with URL) and date of registration.	Page 2 and 13
Protocol and statistical analysis plan	Where the trial protocol and statistical analysis plan can be accessed.	Supplementary
4. Data sharing	Where and how the individual de-identified participant data (including data dictionary), statistical code and any other materials can be accessed.	Page 17
5. Funding and Conflicts of Interest		Page 17-18
5a. Funding	Sources of funding and other support (eg, supply of drugs), and role of funders in the design, conduct, analysis, and reporting of the trial.	Page 17
5b. Conflicts of interest	Financial and other conflicts of interest of the manuscript authors.	Page 18
Introduction		
6. Background and rationale	Scientific background and rationale.	Page 3
7. Objectives	Specific objectives related to benefits and harms.	Page 3
Methods		
8. Patient and public involvement	Details of patient or public involvement in the design, conduct and reporting of the trial.	Page 16
9. Trial Design	Description of trial design including type of trial (eg, parallel group, crossover), allocation ratio, and framework (eg, superiority, equivalence, non-inferiority, exploratory).	Page 13

10. Changes to trial protocol	Important changes to the trial after it commenced including any outcomes or analyses that were not pre-specified, with reason.	Not applicable, no changes.
11. Trial Setting	Settings (eg, community, hospital) and locations (eg, countries, sites) where the trial was conducted.	Page 13
12. Eligibility Criteria		Page 13
12a. Participants	Eligibility criteria for participants.	Page 13
12b. Other	If applicable, eligibility criteria for sites and for individuals delivering the interventions (eg, surgeons, physiotherapists).	Not applicable
13. Intervention and comparator	Intervention and comparator with sufficient details to allow replication. If relevant, where additional materials describing the intervention and comparator (eg, intervention manual) can be accessed.	Page 14
14. Outcomes	Prespecified primary and secondary outcomes, including the specific measurement variable (eg, systolic blood pressure), analysis metric (eg, change from baseline, final value, time to event), method of aggregation (eg, median, proportion), and time point for each outcome.	Page 15
15. Harms	How harms were defined and assessed (eg, systematically, non-systematically).	Page 16
16. Sample Size		Page 15
16a. How sample size was determined	How sample size was determined, including all assumptions supporting the sample size calculation.	Page 15
16b. Interim analyses and stopping criteria	Explanation of any interim analyses and stopping guidelines.	Not applicable, no interim analyses planned.
17. Randomisation		Page 14
17a. Sequence Generation	Who generated the random allocation sequence and the method used.	Page 14

17b. Type of Randomisation	Type of randomisation and details of any restriction (eg, stratification, blocking, and block size).	Page 14
18. Allocation concealment mechanism	Mechanism used to implement the random allocation sequence (eg, central computer/telephone; sequentially numbered, opaque, sealed containers), describing any steps to conceal the sequence until interventions were assigned.	Page 14
19. Implementation	Whether the personnel who enrolled and those who assigned participants to the interventions had access to the random allocation sequence.	Page 14
20. Blinding		Page 14
20a. Who was blinded	Who was blinded after assignment to interventions (eg, participants, care providers, outcome assessors, data analysts).	Page 14
20b. How blinding was achieved	If blinded, how blinding was achieved and description of the similarity of interventions.	Page 14
21. Statistical methods		Page 15-16
21a. Comparing groups	Statistical methods used to compare groups for primary and secondary outcomes, including harms.	Page 16
21b. Definition of who is included in each analysis	Definition of who is included in each analysis (e.g., all randomised participants), and in which group.	Page 16
21c. Missing Data	How missing data were handled in the analysis.	Not applicable, no missing data.
21d. Additional Analyses	Methods for any additional analyses (eg, subgroup and sensitivity analyses), distinguishing pre-specified from post hoc.	Page 16
22. Participant flow, including flow diagram		Figure 2
22a. Participant Numbers	For each group, the numbers of participants who were randomly assigned, received intended intervention, and were analysed for the primary outcome.	Page 5

22b. Losses and exclusions	For each group, losses and exclusions after randomisation, together with reasons.	Page 5
23. Recruitment		Page 5
23a. Dates	Dates defining the periods of recruitment and follow-up for outcomes of benefits and harms.	Page 5
23b. Reasons for stopping recruitment	If relevant, why the trial ended or was stopped.	Not applicable
24. Intervention and comparator delivery		Page 13-14
24a. As Administered	Intervention and comparator as they were actually administered (eg, where appropriate, who delivered the intervention/comparator, whether participants adhered, whether they were delivered as intended (fidelity).	Page 13-14
24b. Concomitant Care	Concomitant care received during the trial for each group.	Page 14
25. Baseline Data	A table showing baseline demographic and clinical characteristics for each group.	Table 1
26. Numbers analysed, outcomes, and estimation	For each primary and secondary outcome, by group: - the number of participants included in the analysis. - the number of participants with available data at the outcome time point. - result for each group, and the estimated effect size and its precision (such as 95% confidence interval). - for binary outcomes, presentation of both absolute and relative effect size.	Page 5; Table 2; Table 3
27. Harms	All harms or unintended events in each group.	Not applicable, no adverse events reported.

28. Ancillary Analyses	Any other analyses performed, including subgroup and sensitivity analyses, distinguishing pre-specified from post hoc.	Page 6
Discussion		
29. Interpretation	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence.	Page 7-10
30. Limitations	Trial limitations, addressing sources of potential bias, imprecision, generalisability, and, if relevant, multiplicity of analyses.	Page 9-10

Trial Protocol

Performance of a Large Language Model Agent vs Ophthalmology Residents in Pre-consultation: A Randomized Clinical Trial

Principle Investigator: Haotian Lin^{1,2,3}

¹ Zhongshan Ophthalmic Center, Sun Yat-sen University, WHO Collaborating Centre for Eye Care and Vision, State Key Laboratory of Ophthalmology, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangdong Provincial Clinical Research Center for Ocular Diseases, Guangzhou, China.

² Hainan Eye Hospital and Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Haikou, China.

³ Department of Genetics and Biomedical Informatics, Zhongshan School of Medicine, Institute for Frontier Interdisciplinary Research in Health Sciences and Technology, Sun Yat-sen University, Guangzhou, China.

Version 1.0 (January 20, 2025): Original protocol.

Structured summary	
Title	Performance of a Large Language Model Agent vs Ophthalmology Residents in Pre-consultation: A Randomized Clinical Trial
Trial Design	A prospective, parallel-group (1:1), open-label, superiority randomized clinical trial.
Primary Outcome	Pre-consultation quality, measured by the MedHistory composite score (range, 0-100).
Secondary Outcomes	Patient satisfaction, pre-consultation duration, and the appropriateness of test recommendation.
Main Trial Hypothesis	Compared with the traditional ophthalmology resident approach, the LLM-based agent is expected to demonstrate superior pre-consultation performance.
Eligibility criteria	Inclusion criteria: 1. Patients of any age or sex. 2. Presentation with non-emergency ocular conditions, including corneal diseases, lens disorders, and vitreoretinal diseases. 3. Willingness to participate with provision of written informed consent. Exclusion criteria: Top 10 ocular emergencies, including globe perforation, ocular chemical injury, corneal ulcer perforation, pseudomonas aeruginosa keratitis, acute angle-closure glaucoma, acute panophthalmitis, central retinal artery occlusion, acute optic neuritis, endophthalmitis, and orbital cellulitis.
Sample size	Target enrollment 172 participants.
Intervention	LLM-based agent vs Residents conduct pre-consultation (history-taking and documentation).
Study setting and locations	A single tertiary ophthalmic hospital in Guangzhou, China.

Section 1: Administrative information.....	51
1.1 Competing interests.....	51
1.2 Funding.....	51
1.3 Sponsor.....	51
1.4 Roles of protocol contributors.....	51
1.5 Data sharing.....	51
Section 2: Introduction.....	52
2.1 Background and rationale.....	52
2.2 Objectives.....	53
Section 3: Study methods.....	54
3.1 Trial design.....	54
3.2 Trial setting.....	54
3.3 Eligibility criteria.....	54
3.4 Intervention and comparator.....	55
3.5 Outcomes.....	56
3.6 Harms.....	59
3.7 Participant timeline.....	60
3.8 Sample size.....	62
3.9 Randomization and masking.....	62
3.10 Data collection methods.....	62
3.11 Data management.....	64
3.12 Statistical methods.....	64
3.13 Monitoring.....	66
Section 4: Ethics.....	66
4.1 Approval of the clinical study protocol and amendments.....	66
4.2 Participant information and informed consent.....	67
4.3 Confidentiality.....	67
4.4 Ancillary and post-trial care.....	67
4.5 Publication policy.....	68
Section 5: Focus group discussion report: Development of history-taking evaluation criteria by ophthalmologists.....	68
5.1 Objective.....	68
5.2 Methodology & Design.....	68
5.3 Summary of focus group discussion.....	69
Section 6: Focus group discussion report: Patient experience and questionnaire development...	72
6.1 Objective.....	72
6.2 Methodology & Design.....	72
6.3 Thematic analysis and discussion rounds.....	73
6.4 Consensus and Item generation.....	74
6.5 Final Instrument.....	75
Section 7: References.....	75

Section 1: Administrative information

1.1 Competing interests

The authors declare that they have no competing interests.

1.2 Funding

This work is supported by the National Natural Science Foundation of China (Grant No. 82441003), the National Natural Science Foundation of China (Grant No. 92368205), the Natural Science Foundation of Guangdong Province in China (Grant No. 2026A1515012947), the Natural Science Foundation of Guangdong Province (Grant No. 2024A1515012292), the State Key Laboratory of Ophthalmology (Grant No. 2025QZLH05), and the Basic scientific research projects of Sun Yat-sen University (Grant No. 23ykcxqt002). The funders are not involved in the study design, data collection, analysis, interpretation, manuscript writing or in the decision to submit the article for publication.

1.3 Sponsor

Zhongshan Ophthalmic Center, Sun Yat-sen University serve as the sponsor and is responsible for study design, conduct, data management, and regulatory compliance.

1.4 Roles of protocol contributors

HTL is the Principal Investigator of the study. HTL, MJL, YXL design the study; MJL, YXL, LJ write the protocol.

1.5 Data sharing

Deidentified data are available upon reasonable request. Researchers interested in accessing the data for methodologically sound research should direct their inquiry to the corresponding authors. Data sharing will be required to complete an institutional Data Use Agreement.

Section 2: Introduction

2.1 Background and rationale

Outpatient healthcare systems are increasingly strained by the disparity between growing patient volumes and available clinical staff^{1,2}. This systemic burden necessitates shorter consultation times, potentially compromising the quality of the physician-patient interaction^{3,4}. As the foundation of clinical assessment, medical history acquisition is particularly vulnerable to these constraints. Whether conducted by physicians, residents, or allied health professionals, high-quality history-taking is a time-consuming and cognitively demanding task that requires synthesizing attentive listening with targeted clinical inquiry⁵. Under current pressures, this process is frequently truncated, leading to incomplete data collection and missed diagnostic cues^{6,7}. To optimize workflow efficiency and care quality, pre-consultation systems have been proposed to acquire and document clinical status prior to the physician encounter.

Conventional pre-consultation workflows typically rely on static questionnaires or rule-based algorithms that present multiple-choice inquiries from a finite, predefined pool⁸. These systems are often constrained by rigid decision rules and a lack of conversational interactivity, limiting their ability to capture complex, unstructured patient narratives. The emergence of large language models (LLMs), with their capacity for context-aware processing, natural language interaction and capability of autonomous task execution, introduces a potential solution⁹. LLMs have demonstrated proficiency in various clinical tasks, including simplifying discharge records¹⁰, summarizing clinical documentation¹¹, and ambiently transcribing consultations¹². However, these applications predominantly deploy LLMs as reactive information processors. Whether these models can function as active clinical agents, capable of initiating and guiding appropriate inquiry to conduct clinically meaningful history-taking, remains underexplored.

To enhance the clinical utility of these models, the chain-of-thought (CoT) reasoning paradigm has been adopted to facilitate complex, multistep medical decision-making^{13,14}. Studies employing CoT have reported that LLMs can perform comparably to primary care physicians in history-taking tasks¹⁵. However, prior evaluations have largely been confined to retrospective or simulated data using general-purpose models, rather than specialized agents integrated into real-world clinical workflows. This limits their validity regarding actual patient dynamics and clinic flow integration. Furthermore, rigorous clinical evaluation of open-ended tasks remains challenging¹⁶. While prior efforts have adapted established frameworks, such as the Objective Structured Clinical Examination (OSCE)¹⁷, these metrics typically assess generalist competencies. Consequently, they lack the domain-specific granularity required to evaluate the nuanced diagnostic interrogation essential for specialty care.

To systematically address these challenges, we developed a specialized ophthalmology LLM agent for pre-consultation, together with a multidimensional evaluation framework (MedHistory). This framework was grounded in consensus from focus group discussions and adapted from the OSCE¹⁸. Building on our previously validated methodologies¹⁹, the primary objective of this study is to evaluate the effectiveness of the LLM-based pre-consultation agent compared with ophthalmology residents within a real-world outpatient workflow in a prospective randomized clinical trial (RCT).

2.2 Objectives

Primary objective:

To assess the effect of the LLM-based agent on improving the quality of pre-consultation tasks in real-world outpatient clinical setting.

Hypothesis related to the primary objective:

The LLM-based agent will achieve superior pre-consultation performance, as reflected by a higher composite MedHistory score.

Secondary objectives:

To compare patient satisfaction, pre-consultation duration, and the appropriateness of test recommendation between groups.

Hypotheses related to the secondary objectives:

The LLM-based agent will require a longer pre-consultation duration while maintaining or improving patient satisfaction and test recommendation performance.

Section 3: Study methods**3.1 Trial design**

A single-center, prospective, parallel-group (1:1), open-label, superiority randomized clinical trial.

3.2 Trial setting

Zhongshan Ophthalmic Center, Sun Yat-sen University—a single tertiary ophthalmic hospital in Guangzhou, China.

3.3 Eligibility criteria**Inclusion criteria:**

- (1) Patients of any age or sex.
- (2) Presentation with non-emergency ocular conditions, including corneal diseases, lens disorders, and vitreoretinal diseases.
- (3) Willingness to participate with provision of written informed consent.

Exclusion criteria:

Top 10 ocular emergencies, including globe perforation, ocular chemical injury, corneal ulcer perforation, pseudomonas aeruginosa keratitis, acute angle-closure glaucoma, acute panophthalmitis, central retinal artery occlusion, acute optic neuritis, endophthalmitis, and orbital cellulitis.

3.4 Intervention and comparator

Patients registering for a non-emergency appointment at the clinic of Zhongshan Ophthalmic Center will be approached and consecutively screened by clinical assistants regarding their visit purpose while waiting in the lobby. Participants expressing interest in the study will receive full disclosure and will be clearly informed that the study should not be construed as providing direct clinical guidance. Individuals are enrolled according to the study eligibility criteria after providing written informed consent. Participants could receive assistance if needed.

LLM arm: Patients will access the conversational agent via the hospital's mobile portal. Following online registration, a demonstration of the agent pre-consultation process was provided. Interactions will be conducted via audio or text, with all exchanges transcribed as text logs. Upon completion, the agent will generate a structured clinical chart that will be automatically synchronized with the Hospital Information System (HIS).

Traditional arm: Residents will conduct history-taking and document the encounter directly in the HIS. All interactions will be audio-recorded and subsequently transcribed. All residents have completed the hospital's pre-service training, and no standardized documentation template will be used.

Following the pre-consultation, patients in both groups will complete anonymous satisfaction questionnaires before proceeding to clinical consultants. The consultants will review the preliminary charts (generated by either the agent or the resident), perform slit-lamp examinations, and make necessary amendments. The finalized documentation, encompassing patient history, test recommendations, primary diagnosis, and treatment plan, will serve as the reference standard for the study.

Trained clinical assistants, familiar with routine outpatient workflows and the trial procedures, will be available to provide technical assistance if needed, without influencing clinical interactions. In cases of participant withdrawal or clinical deterioration, the allocated intervention will be discontinued, and routine clinical care will be provided. No additional AI-based systems are permitted during the study encounter.

3.5 Outcomes

Primary outcome: pre-consultation quality

The primary outcome is the pre-consultation quality, measured by the MedHistory composite score (range, 0-100), which evaluates history-taking, communication and documentation. Subscale scores are reported as component measures. The MedHistory framework was grounded in consensus from focus group discussions (Section 5) and adapted from the OSCE¹⁸. Data are aggregated as mean $\pm$ SD and assessed by two independent evaluators upon study completion up to one week.

Secondary outcomes

Secondary outcome 1: patient satisfaction

Patient satisfaction is measured using an eight-item questionnaire rated on a 5-point Likert scale (1–5: very dissatisfied–very satisfied), developed based on patient focus group discussions (Section 6). Scores are

summarized as the median (IQR). The questionnaire is administered during the outpatient visit and reviewed within one week after study completion.

Secondary outcome 2: pre-consultation duration

Pre-consultation duration is defined as the time from the initiation of history-taking to the completion of documentation, measured in minutes, aggregated as median [interquartile range (IQR)]. The duration will be recorded during the outpatient visit and reviewed upon study completion up to one week.

Secondary outcome 3: appropriateness of test recommendation

The performance of test recommendation based on the information of pre-consultation will be evaluated using a predefined expert rating scale (0–3: completely wrong–fully appropriate) and quantitatively via recall, precision and F1 score against the reference standard established by senior consultant. Qualitative ratings are summarized as median (IQR) and distribution across score categories frequency (percentage); quantitative metrics are calculated using micro-level aggregation. Test recommendations are collected during the outpatient visit and will be reviewed upon study completion up to one week.

Test recommendation performance metrics (recall, precision, and F1 score) are focused exclusively on mandatory examinations (defined as tests essential for identifying the principal clinical condition), reflecting the model's ability to identify the principal clinical condition driving management decisions. Secondary or ancillary tests are not included due to their variable clinical relevance and the lack of a standardized reference boundaries.

Exploratory outcomes: signs-integration stage

Routine outpatient encounters involve history-taking followed by assessment of ocular signs. Exploratory outcomes will therefore be assessed at the Signs-Integration stage, in which textual descriptions of ocular signs, including slit-lamp findings, visual acuity, and intraocular pressure, are provided after the initial history-only stage.

Exploratory outcome 1: primary diagnosis performance

Diagnostic performance will be evaluated qualitatively using a predefined expert rating scale (0–3: completely incorrect–optimal performance) and quantitatively via recall, precision, and F1 score against the reference standard established by senior consultant. Qualitative ratings are summarized as median (IQR) and distribution across score categories number (percentage); quantitative metrics are reported using macro-level and micro-level aggregation. The diagnostic outputs are generated during the outpatient visit and will be reviewed upon study completion up to one week.

Diagnostic performance metrics (recall, precision, and F1 score) are calculated for primary diagnoses only, reflecting the model’s ability to identify the principal clinical condition driving management decisions. Secondary or ancillary diagnoses are not included in quantitative diagnostic performance evaluation due to their variable clinical relevance and lack of a standardized reference boundary.

Exploratory outcome 2: treatment plan formulation quality

Treatment plan formulation quality will be evaluated using a predefined expert rating scale (0–3: completely incorrect–principally correct). Scores are summarized as median (IQR) and distribution across score categories number (percentage). The treatment plans are generated during the outpatient visit and will be reviewed upon study completion up to one week.

Exploratory outcome 3: projected test-related expenses and wait times

Projected test-related expenses and wait times will be estimated based on institutional data for the recommended tests, with wait time defined as the maximum duration among recommended tests.

Measurement variables include total projected cost and projected wait time per participant, aggregated as median (IQR). These projected outcomes will be derived and reviewed upon study completion up to one week.

3.6 Harms

Given the non-invasive and non-therapeutic nature of the intervention, harms are expected to be minimal.

Harms are defined as any adverse events or unintended negative effects related to the pre-consultation process, including patient distress, misunderstanding of medical information, delays in clinical care, or workflow disruptions. Harms will be assessed through spontaneous reporting of any concerns by participants, clinicians, or medical staff during the study encounter. All reported harms will be documented and reviewed by the study investigators, and appropriate clinical care will be provided as needed. For adverse events, investigators will analyze the causes, conduct follow-up observations, and maintain proper records. All adverse events occurring during the study will be documented, and the severity of each adverse event will be assessed using the Common Terminology Criteria for Adverse Events (CTCAE v5.0). The case report form (CRF) will record the occurrence date, duration, severity, management, and outcome of each adverse event. All adverse events will be followed up until they resolve or until the investigator determines they are clear.

3.7 Participant timeline

The study follows a single-visit outpatient design. After enrollment and randomization, participants undergo the allocated pre-consultation process (LLM-based agent or resident-led), followed by routine clinical care.

No run-in or washout periods are applied.

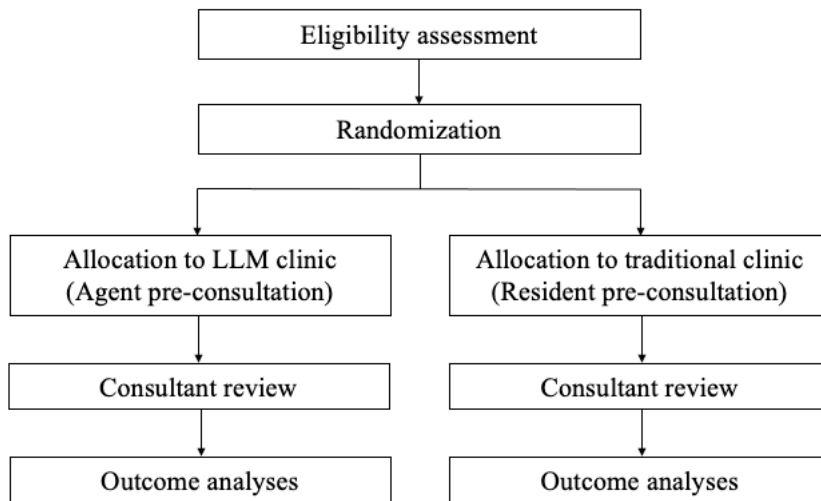


Figure 1. Schematic diagram of the study design

Table1. Time schedule of enrollment, interventions, and assessments for participants.

Study period	Enrollment	Allocation	Pre-consultation	Consultant review
Time point	Day 0	Day 0	Day 0	Day 0
Informed consent	✓			
Eligibility screening	✓			
Randomization		✓		
History-taking and documentation (LLM-based agent or resident-led)			✓	
Patient satisfaction questionnaire			✓	
Consultant review and slit-lamp examination				✓
Finalized clinical documentation (reference standard)				✓

3.8 Sample size

The sample size of 172 participants (86 per group) was calculated in PASS 16 (NCSS, USA) based on the primary outcome to achieve 90% power with a two-sided α of 0.05, assuming an 8-point between-group difference and a standard deviation of 16 derived from an unpublished 30-participant pilot study.

The pilot study was conducted on December 16, 2024, in the same outpatient clinical setting to inform sample size estimation. It followed the planned randomized controlled trial workflow, including identical eligibility criteria, clinical procedures, and outcome assessments.

3.9 Randomization and masking

The randomization sequence will be generated by an independent statistician using a computer-based block randomization procedure implemented in R (version 4.0.3). Sequentially numbered, opaque, sealed envelopes will be prepared in advance according to the randomization list to ensure allocation concealment. After providing consent, participants will be informed of their allocation by the clinical assistants.

Participants and the clinical assistants will be aware of the group assignment due to the intervention nature.

Participants will not be informed of the trial hypotheses. The ophthalmology residents (with ≥ 3 years of experience) will only be aware of participants allocated to the traditional clinic and will remain masked to the study hypotheses. The clinical consultants and two independent evaluators (all with > 5 years of ophthalmology experience), and the statistician will be masked to the study hypotheses and group allocation.

3.10 Data collection methods

Pre-consultation processes, duration, test recommendations, primary diagnoses, and treatment plans in the intervention group are collected through the LLM-based agent system. In the control group, the pre-

consultation process and duration are audio-recorded and subsequently transcribed; documentation, test recommendations, primary diagnoses and treatment plans are extracted from the hospital information system (HIS). Final consultant-reviewed clinical documentation is also obtained from the HIS. Baseline characteristics and patient questionnaires are collected using paper-based case report forms (CRFs).

Assessment instruments

The MedHistory framework, developed by an ophthalmic specialist focus group, is used to assess the quality of pre-consultation history-taking, communication, and documentation. The quality of test recommendations, diagnostic performance, and treatment plan formulation is evaluated using predefined expert rating scales (0–3) developed by senior ophthalmologists. Patient questionnaires are developed by the study investigators. All outcomes are independently assessed by two evaluators with more than five years of clinical experience in ophthalmology. Inter-rater agreement will be assessed and reported where applicable. The same reference standard was consistently applied across all assessment phases within each study group.

The quantitative assessment of test recommendations and diagnostic performance: For each case i , test recommendation or primary diagnostic performance is evaluated using a set-based framework. Let T_i denote the reference standard set established by specialists, and C_i denote the set of recommended tests or predicted diagnoses generated during pre-consultation in the LLM-based agent and resident groups. **True positives** are defined as items present in both C_i and T_i ; **false positives** as items present in C_i but not included in T_i ; and **false negatives** as items present in T_i but not predicted or recommended in C_i .

Recall is defined as the proportion of items correctly identified in the reference standard ($TP_i / |T_i|$).

Precision is defined as the proportion of recommended test or predicted diagnoses that are correct ($TP_i / |C_i|$).

F1 score is defined as the harmonic mean of recall and precision. When both precision and recall are zero, the F1 score is defined as zero.

Accuracy is not computed because true negatives are not well defined in this test recommendation or diagnostic setting. The space of potentially non-recommended tests or non-diagnosed conditions is large and clinically unbounded, and defining true negatives would require arbitrary assumptions regarding irrelevant tests or conditions. As a result, accuracy would be dominated by true negatives and potentially misleading. Accordingly, recall, precision, and F1 score are selected as the primary performance metrics, as they more appropriately capture omission of required tests or missed diagnoses (false negatives), recommendation of inappropriate tests or incorrect diagnoses (false positives), and the balance between these two aspects of clinical decision-making quality.

3.11 Data management

Data will be de-identified using a standardized anonymization process to remove all Protected Health Information (PHI), adhering to the Health Insurance Portability and Accountability Act (HIPAA). Each case will be assigned a unique study ID. Interaction dialogue, clinical charts, test recommendations, primary diagnoses, and treatment plans for both groups will be compiled into standardized spreadsheets for evaluation.

3.12 Statistical methods

All randomized participants will be included in the analysis.

Level of significance

Statistical significance is defined as a two-sided $P < .05$ for all analyses.

Handling of missing data

Multiple imputation.

Primary outcome: pre-consultation quality

The primary outcome, the MedHistory composite score, will be analyzed using multivariable linear regression, adjusting for age and sex; subscale scores (history-taking, communication, and documentation quality) are reported as component measures.

Secondary outcome 1: patient satisfaction

Patient satisfaction scores will be compared between groups using the Mann–Whitney U test.

Secondary outcome 2: pre-consultation duration

Pre-consultation duration will be compared between groups using the Mann–Whitney U test between groups.

Secondary outcome 3: appropriateness of test recommendation

Qualitative performance will be compared between groups using the Mann–Whitney U test. Quantitative performance metrics (recall, precision, F1 score) will be calculated at both macrolevel and microlevel, with 95% confidence intervals (CIs) and two-sided P values estimated using bootstrap resampling (2000 iterations).

Exploratory outcome 1: primary diagnosis performance

Quantitative performance metrics (recall, precision, F1 score) will be calculated at both macrolevel and microlevel, with 95% confidence intervals (CIs) and two-sided P values estimated using bootstrap resampling (2000 iterations).

To evaluate performance differences across stages, linear mixed-effects models will be fitted with fixed effects for group (agent vs resident), stage (signs-integration vs history-only), and their interaction, with a participant-level random intercept, adjusting for age and sex.

Exploratory outcome 2: treatment plan formulation quality

Qualitative performance will be compared between groups using the Mann–Whitney U test.

Exploratory outcome 3: projected test-related expenses and wait times

Group differences will be assessed using the Mann–Whitney U tests, with Hodges–Lehmann estimates reported as measures of effect size.

3.13 Monitoring

The data will be subject to independent monitoring and quality auditing by the Clinical Research Center (CRC) throughout the conduct of the trial.

Given the low-risk nature of the study and its conduct within routine outpatient clinical workflows, no external trial monitoring is planned. Trial conduct is overseen on-site by trained clinical assistants who are familiar with routine outpatient procedures and study-specific workflows. The clinical assistants are responsible for coordinating participant flow, ensuring adherence to the study protocol, and providing technical support as needed, without influencing clinical interactions.

Section 4: Ethics

4.1 Approval of the clinical study protocol and amendments

Before the study begins, the study protocol, protocol amendments, informed consent form, investigator's brochure, and other relevant documents will be submitted to the ethics committee for review and approval.

Major amendments to the study protocol will be submitted for ethics committee approval prior to implementation, except for changes made to promptly eliminate immediate hazards to participants or for study management purposes. If the investigator modifies or deviates from the protocol without prior ethics committee approval to eliminate an immediate hazard to participants, the investigator will promptly report the changes to the ethics committee and provide justification.

4.2 Participant information and informed consent

Researchers will fully inform participants about the details of the study, including its potential benefits and risks, and will make every effort to answer all questions raised by the participants regarding the study.

Participants will be informed that participation in the study is voluntary. Before participating, participants will sign an ethics committee–approved informed consent form. If researchers obtain new information that may affect a participant’s continued participation in the study, the participant will be promptly informed, and the update will be properly documented. Participation is entirely voluntary, and refusal to participate will not affect the medical care received by the participant. Participants may withdraw from the study at any time.

4.3 Confidentiality

All participant information will be handled confidentially throughout the study. Personal identifiers will not be included in analytic datasets, and study data will be accessed only by authorized study personnel. Data will be stored on secure, access-controlled systems in accordance with institutional policies. Any data shared outside the study team will be fully de-identified.

4.4 Ancillary and post-trial care

This study involves a minimal-risk intervention conducted within routine outpatient clinical care and does not require additional ancillary or post-trial care beyond standard clinical management. Participants will continue

to receive usual medical care regardless of study participation or withdrawal. No financial compensation is provided for participation.

4.5 Publication policy

Public disclosure includes publication of a paper in a scientific journal, abstract submission with a poster or oral presentation at a scientific meeting, or disclosure by other means. No confidential information will be disclosed.

Section 5: Focus group discussion report: Development of history-taking evaluation criteria by ophthalmologists

Date: December 30, 2024

5.1 Objective

The primary objective of this focus group discussion was to establish and finalize a standardized evaluation rubric for comparing patient history collection conducted by a large language model (LLM)-assisted system versus traditional human doctor-led history-taking in an outpatient ophthalmology setting. The discussion aimed to provide professional advice on the specific weighting of comprehensiveness, accuracy, efficiency, and clinical usability to inform the final scoring criteria for the randomized clinical trial.

5.2 Methodology & Design

Participants:

- A panel of 8 medical doctors (4 senior and 4 junior ophthalmologists) from outpatient clinical settings participated to ensure diverse professional perspectives.

Setting:

- The session was conducted in a private conference room to facilitate detailed professional discourse.
- The session was moderated by a neutral facilitator experienced in clinical process optimization.

Procedure:

- **Introduction:** The facilitator outlined the context of the study (LLM vs. Human efficacy).
- **Structured discussion:** Ophthalmologists deliberated on key dimensions of history taking, specifically discussing the clinical importance (weighting) of specific history components.
- **Consensus formation:** The group iteratively adjusted scoring weights until a unanimous agreement was reached on the final rubric.
- **Data capture:** The session was audio-recorded, transcribed, and analyzed to extract the final criteria.

5.3 Summary of focus group discussion

The focus group analyzed the requirements for high-quality ophthalmic history taking and agreed upon a 100-point total scale. The consensus discussions that defined the specific criteria are summarized below:

A. History-Taking Content (92 Points Total)

The group determined that the core accuracy of medical information should account for the vast majority of the score (92%). This high weighting reflects the consensus that history taking is a knowledge-intensive process that integrates extensive diagnostic reasoning with strategic inquiry. Since the effective elicitation of clinical facts represents the core value of the medical consultation, the rubric prioritizes the completeness and precision of the information obtained.

I. Chief Complaint (30 Points)

- **Consensus:** Ophthalmologists agreed that the Chief Complaint is the cornerstone of the diagnosis. Missing or vague information here is critical.

- **Breakdown:** To ensure precision, the 30 points were split evenly (10 points each) among three non-negotiable elements:

- o *Symptoms:* Must be clear and specific (e.g., sudden onset).
- o *Duration:* Precise timelines are required for assessment.
- o *Laterality:* In ophthalmology, specifying the eye (left/right/both) is mandatory.

II. History of Present Illness (30 Points)

- **Consensus:** The narrative of the illness requires detailed structuring.
- **Breakdown:**
 - o *Precipitating Factors (10 points):* High weight was assigned here because identifying triggers (e.g., trauma, chemical exposure) fundamentally changes the management plan.
 - o *Characteristics (10 points):* The group emphasized that a complete history must capture both positive symptoms (what is there) and relevant negative symptoms (what is absent, for differential diagnosis).
 - o *Clinical Course (5 points):* This was subdivided to ensure the LLM captures the timeline (1), provisional diagnoses (1), and response to previous treatments (3).
 - o *Progression (5 points):* Evaluating whether the condition is worsening or stable.

III. Past Medical History (15 Points)

- **Consensus:** The group argued that generic Past Medical History is insufficient for eye care. Specific ophthalmic history is required.
- **Breakdown:**
 - o *Spectacle/Contact Lens History (1 point):* Essential for refractive contexts.

- o *General Health (4 points)*: Focusing on systemic diseases that affect the eyes (e.g., diabetes).
- o *Trauma History (5 points)*: Given its impact on structural eye integrity, trauma history was weighted heavily.
- o *Ocular Surgical History (5 points)*: Previous eye surgeries define current anatomical expectations.

IV. Secondary History Components (15 Points Total)

- **Family History (5 points)**: Essential for hereditary conditions (e.g., glaucoma).
- **Allergy History (5 points)**: Critical for patient safety regarding drug prescriptions.
- **Personal History (5 points)**: Lifestyle factors (eye-use habits, occupation, smoking) that impact eye health.

V. Logic Sequencing (2 Points)

- **Consensus**: Beyond the facts, the order of questions matters.
- **Breakdown**: A small but distinct score was assigned to ensure the history is obtained in a coherent, medical logic, rather than a random scattering of questions.

B. Soft Skills and Documentation (8 Points Total)

The group agreed that while clinical facts are paramount, the manner of interaction and the quality of the record are vital for clinical adoption.

I. Communication (4 Points)

- **Empathy (2 points)**: To assess patient-centered language.
- **Clarity (2 points)**: To ensure the patient understood the questions without needing repetition.

II. Documentation (4 Points)

- **Terminology (2 points):** The interviewer must convert colloquial patient speech into formal medical terminology.
- **Consistency (2 points):** The final record must not contradict the patient's spoken statements.

Section 6: Focus group discussion report: Patient experience and questionnaire development

Date: December 31, 2024

6.1 Objective

The objective of this focus group series was to gather qualitative insights into patient satisfaction and experiences regarding history-taking processes. Specifically, the discussion aimed to identify key psychological and logistical factors—such as anxiety, communication clarity, and trust—to assist in developing a validated Patient Experience Questionnaire (secondary outcome) for the randomized clinical trial comparing LLM-assisted systems vs ophthalmology residents.

6.2 Methodology & Design

Participants:

- Patients recruited from the outpatient clinic waiting area.
- **Sample Size:** 12 participants total, divided into two distinct groups (Round 1 and Round 2) of 6 participants each to ensure validation of themes.

Setting:

- Conducted in a private, quiet room to ensure confidentiality and encourage open dialogue.
- Moderated by an independent facilitator neutral to the study design.

Procedure:

- **Introduction:** Facilitator explained the purpose (improving the consultation process).
- **Structured discussion:** Open-ended questions regarding current dissatisfaction with doctor-led history taking and hypothetical acceptance of AI.
- **Consensus:** Identification of specific metrics that matter most to patients.

6.3 Thematic analysis and discussion rounds

Round 1: Identifying core pain points (exploratory phase)

Based on discussion from Group 1

The first group focused on their immediate emotional state during standard consultations. Key themes emerged regarding the "rushed" nature of clinical visits.

- **Anxiety and Nervousness:** Participants explicitly stated they felt "nervous" and afraid they would "forget symptoms" or "couldn't catch the doctor's words."
- **Lack of Explanations:** A recurring complaint was that doctors "fired questions" and issued test orders without addressing the patient's immediate concerns or explaining why.
- **The Role of AI:** Participants agreed that an LLM pre-consultation would "settle nerves" by allowing them to organize their thoughts without time pressure.
- **Privacy:** Concerns were raised regarding data encryption, establishing "Trust" as a necessary metric.

Round 2: Validating themes and refining metrics (confirmatory phase)

Based on discussion from Group 2

The second group validated the themes from Round 1 and helped refine the language for the questionnaire items.

- **Patience vs. Efficiency:** Participants noted that while doctors were polite, they lacked patience ("rushing to the next act," "stared at the screen"). This confirmed the need to measure "Patience" and "Natural Conversation" distinct from simple "Politeness."
- **Empathy:** Participants described interactions as "procedural courtesy" rather than genuine care ("didn't feel he understood my anxiety"). This led to the inclusion of a specific "Empathy" metric.
- **Acceptance:** The group debated the "AI-first" model. While some elderly participants worried about usability ("wrong tap"), the consensus was a willingness to use it as an alternative or supplement, provided privacy was assured.

6.4 Consensus and Item generation

Based on the qualitative data from both rounds, the researchers derived specific survey items to measure the patient experience quantitatively.

Qualitative Theme / Patient Quote	Derived Questionnaire Item
"I was nervous... afraid I'd forget symptoms." / "Telling AI first would settle my nerves."	The agent/resident helped reduce my anxiety.
"Fired questions so fast." / "Rushed to the next act." / "Did not repeat questions."	The agent/resident showed good patience.
"It felt like filling out a form." / "Stared at the screen... official... no warmth."	The conversation felt relaxed and natural.
"No—everything followed the preset protocol." / "I felt my key information wasn't heard."	The agent/resident addressed my concerns before asking follow-up questions.
"Procedural courtesy." / "Didn't feel he understood my anxiety."	The agent/resident showed empathy.
"I would support it." / "Ideally I could keep asking AI questions after I get home."	I am willing to use AI pre-consultation as an alternative to physician-led pre-consultation.

<i>"As long as encryption is like the hospital's." /</i>	I trusted that my privacy would be protected during
<i>"Data security worries me."</i>	the AI pre-consultation.
<i>General consensus on the quality of the interaction</i>	Overall, I was satisfied with the medical service
<i>and outcome.</i>	provided by agent/resident.

6.5 Final Instrument

The focus group findings were synthesized into the following 8-item questionnaire. This instrument will serve as the secondary outcome measure for the study.

Instructions: Please rate your agreement with the following statements regarding your consultation experience today.

Scale: 5-point Likert scale (1: Very Dissatisfied / Strongly Disagree — 5: Very Satisfied / Strongly Agree)

- (1). The agent/resident helped reduce my anxiety.
- (2). The agent/resident showed good patience.
- (3). The conversation felt relaxed and natural.
- (4). The agent/resident addressed my concerns before asking follow-up questions.
- (5). The agent/resident showed empathy.
- (6). I am willing to use AI pre-consultation as an alternative to physician-led pre-consultation.
- (7). I trusted that my privacy would be protected during the AI pre-consultation.
- (8). Overall, I was satisfied with the medical service provided by agent/resident.

Section 7: References

- 1 Rupert Bourne, GBD 2019 Blindness and Vision Impairment Collaborators, Vision Loss Expert Group of the Global Burden of Disease Study. Trends in prevalence of blindness and distance and near vision

- impairment over 30 years: an analysis for the Global Burden of Disease Study. *Lancet Glob Health* 2021; **9**: e130–43.
- 2 Siegel H, Böhringer D, Wacker K, Josef Lopez Niedenhoff P, Mittelviefhaus H, Reinhard T. Duration of Consultations in an Outpatient Ophthalmology Unit. *Dtsch Ärztebl Int* 2023; **120**: 481–2.
 - 3 Baxter SL, Gali HE, Huang AE, *et al*. Time requirements of paper-based clinical workflows and after-hours documentation in a multi-specialty academic ophthalmology practice. *Am J Ophthalmol* 2019; S0002-9394(19)30120-5.
 - 4 Bao Y, Fan G, Zou D, Wang T, Xue D. Patient experience with outpatient encounters at public hospitals in Shanghai: Examining different aspects of physician services and implications of overcrowding. *PLoS ONE* 2017; **12**: e0171684.
 - 5 Gibelli F, Bailo P, Pesel G, Ricci G. Preserving Patient Stories: Bioethical and Legal Implications Related to the Shift from Traditional to Digital Anamnesis. *Clin Pract* 2024; **14**: 1196–213.
 - 6 Peterson MC, Holbrook JH, Von Hales D, Smith NL, Staker LV. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. *West J Med* 1992; **156**: 163–5.
 - 7 Levine D. History taking is a complex skill. *BMJ* 2017; **358**: j3513.
 - 8 Zhakhina G, Tapinova K, Kanabekova P, Kainazarov T. Pre-consultation history taking systems and their impact on modern practices: Advantages and limitations. *J Clin Med Kazakhstan* 2023; **20**: 26–35.
 - 9 Young BK, Zhao PY. Large Language Models and the Shoreline of Ophthalmology. *JAMA Ophthalmol* 2024; **142**: 375–6.
 - 10 Zaretsky J, Kim JM, Baskharoun S, *et al*. Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format. *JAMA Netw Open* 2024; **7**: e240357.

- 11 Van Veen D, Van Uden C, Blankemeier L, *et al.* Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med* 2024; **30**: 1134–42.
- 12 Irfan AA, Khatim NA, Arief MM. Using LLM for Real-Time Transcription and Summarization of Doctor-Patient Interactions into ePuskesmas in Indonesia. 2024; published online Sept 25. DOI:10.48550/arXiv.2409.17054.
- 13 Singhal K, Tu T, Gottweis J, *et al.* Toward expert-level medical question answering with large language models. *Nat Med* 2025; **31**: 943–50.
- 14 Zhang Z, Yao Y, Zhang A, *et al.* Igniting Language Intelligence: The Hitchhiker’s Guide From Chain-of-Thought Reasoning to Language Agents. 2023; published online Nov 20. <http://arxiv.org/abs/2311.11797> (accessed July 2, 2024).
- 15 Leybold T, Lingens LF, Beier JP, Boos AM. Integrating AI in Lipedema Management: Assessing the Efficacy of GPT-4 as a Consultation Assistant. *Life* 2024; **14**: 646.
- 16 Johri S, Jeong J, Tran BA, *et al.* An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat Med* 2025; **31**: 77–86.
- 17 Huang T-Y, Hsieh PH, Chang Y-C. Performance Comparison of Junior Residents and ChatGPT in the Objective Structured Clinical Examination (OSCE) for Medical History Taking and Documentation of Medical Records: Development and Usability Study. *JMIR Med Educ* 2024; **10**: e59902–e59902.
- 18 Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: An historical and theoretical perspective. *Med Teach* 2013; **35**: e1437–46.
- 19 Luo M-J, Pang J, Bi S, *et al.* Development and Evaluation of a Retrieval-Augmented Large Language Model Framework for Ophthalmology. *JAMA Ophthalmol* 2024; **142**: 798.

Statistical Analysis Plan

SAP Version: 1.0 (January 20, 2025)

Study protocol title: Performance of a Large Language Model Agent vs Ophthalmology Residents in Pre-consultation: A Randomized Clinical Trial

Protocol Version: 1.0 (January 20, 2025)

1. Introduction.....	80
2. Background and rationale.....	80
3. Objectives.....	80
4. Outcome measures	81
4.1 Primary outcome measures	81
4.2 Secondary outcome measures	81
4.3 Exploratory outcome measures	82
5. Study methods	83
5.1 Trial design.....	83
5.2 Randomization and masking	83
5.3 Sample size	83
5.4 Hypotheses.....	84
5.5 Statistical interim analysis	84
5.6 Time of outcome assessments	84
6. Statistical principles.....	84
6.1 Confidence interval and P values	84
6.2 Analysis population	85
6.3 Descriptive statistics	85
7. Statistical methods	85
7.1 Demographic and baseline characteristics	85
7.2 Inter-rater reliability	85
7.3 Primary efficacy analyses.....	85
7.4 Secondary efficacy analyses	86
7.5 Exploratory efficacy analyses	87
7.6 Handling of missing data.....	88
7.7 Subgroup analysis.....	89
7.8 Sensitivity analysis.....	89
7.9 Statistical software.....	89
7.10 Deviations from the SAP	89
8. References	89

1. Introduction

The purpose of this statistical analysis plan (SAP) is to describe the procedures and the statistical methods that will be used to analyse and report results.

2. Background and rationale

Outpatient healthcare systems are increasingly strained by the disparity between growing patient volumes and available clinical staff^{1,2}. As the foundation of clinical assessment, medical history acquisition is particularly vulnerable to these constraints, risking incomplete data collection and missed diagnostic cues^{3,4}.

To optimize workflow efficiency and care quality, pre-consultation systems have been proposed to acquire and document clinical status prior to the physician encounter.

Conventional pre-consultation tools typically rely on static questionnaires or rule-based algorithms, limiting their ability to capture patient narratives⁵. The emergence of large language models (LLMs), with their capacity for context-aware processing and natural language interaction, introduces a potential solution⁶. It has been reported that adapted LLMs can perform comparably to primary care physicians in history-taking tasks⁷. However, prior evaluations have largely been confined to retrospective or simulated data using general-purpose models, rather than specialized agents integrated into real-world clinical workflows. This limits their validity regarding actual patient dynamics and clinic flow integration. Whether LLMs can function effectively as autonomous agents for pre-consultation history-taking, compared with trained clinicians in real-world clinical workflows, remains unproven.

3. Objectives

The primary objective of this trial is to evaluate whether an LLM-based agent improves pre-consultation quality, as measured by the MedHistory composite score, compared with ophthalmology residents.

Secondary objectives are to compare group differences in (1) patient satisfaction (2) pre-consultation duration, and (3) the appropriateness of test recommendations.

Exploratory objectives include evaluating group differences in (1) primary diagnosis performance across history-only and signs-integrated stages, (2) treatment plan formulation quality, and (3) projected test-related expenses and wait times based on recommended tests.

4. Outcome measures

4.1 Primary outcome measures

The primary outcome is the MedHistory composite score, a continuous measure reflecting the quality of pre-consultation history-taking, with subscale scores for history-taking, communication, and documentation quality reported as component measures, all independently assessed by two evaluators.

4.2 Secondary outcome measures

Secondary outcome 1: patient satisfaction

Patient satisfaction is measured using an eight-item questionnaire rated on a 5-point Likert scale (1–5: very dissatisfied–very satisfied).

Secondary outcome 2: pre-consultation duration

Pre-consultation duration is defined as the time from the initiation of history-taking to the completion of documentation, measured in minutes.

Secondary outcome 3: appropriateness of test recommendation

The performance of test recommendation based on the information of pre-consultation is evaluated using a predefined expert rating scale (0–3: completely wrong–fully appropriate) and quantitatively via recall, precision and F1 score against the reference standard established by senior consultant. Quantitative metrics are reported using micro-level aggregation.

4.3 Exploratory outcome measures

Exploratory outcome 1: primary diagnosis performance

Diagnostic performance is evaluated qualitatively using a predefined expert rating scale (0–3: completely incorrect–optimal performance) and quantitatively via recall, precision, and F1 score against the reference standard established by senior consultant. Quantitative metrics are reported using macro-level and micro-level aggregation.

Diagnostic performance metrics (recall, precision, and F1 score) are calculated for primary diagnoses only, reflecting the model’s ability to identify the principal clinical condition driving management decisions.

Secondary or ancillary diagnoses are not included in quantitative diagnostic performance evaluation due to their variable clinical relevance and lack of a standardized reference boundary.

Exploratory outcome 2: treatment plan formulation quality

Treatment plan formulation quality is evaluated using a predefined expert rating scale (0–3: completely incorrect–principally correct).

Exploratory outcome 3: projected test-related expenses and wait times

Projected test-related expenses and wait times are estimated based on institutional data for the recommended tests, with wait time defined as the maximum duration among recommended tests. Measurement variables include total projected cost and projected wait time per participant.

5. Study methods

5.1 Trial design

This is a single-center, prospective, parallel-group (1:1), open-label, superiority randomized clinical trial.

5.2 Randomization and masking

The randomization sequence is generated by an independent statistician using a computer-based block randomization procedure implemented in R (version 4.0.3). Sequentially numbered, opaque, sealed envelopes are prepared in advance according to the randomization list to ensure allocation concealment. After providing consent, participants are informed of their allocation by the clinical assistants.

Participants and the clinical assistants are aware of the group assignment due to the intervention nature.

Participants are not informed of the trial hypotheses. The ophthalmology residents (with ≥ 3 years of experience) are only aware of participants allocated to the traditional clinic and remain masked to the study hypotheses. The clinical consultants and two independent evaluators (all with > 5 years of ophthalmology experience), and the statistician are masked to the study hypotheses and group allocation.

5.3 Sample size

The sample size is determined based on the primary outcome of the trial. Assuming a two-sided significance level (α) of 0.05 and a statistical power of 90%, the trial is powered to detect an 8-point between-group

difference, with a common standard deviation of 16 points. These assumptions are informed by an unpublished pilot study involving 30 participants.

Under these assumptions, a total sample size of 172 participants (86 per group) is required. Sample size calculations are performed using PASS version 16 (NCSS, LLC, USA).

The pilot study was conducted on December 16, 2024, in the same outpatient clinical setting as the planned randomized controlled trial and followed an identical workflow, including eligibility criteria, clinical procedures, and outcome assessments, to ensure the applicability of the variance estimates used for sample size determination.

5.4 Hypotheses

The LLM-based agent achieves superior pre-consultation performance, as reflected by a higher composite MedHistory score.

5.5 Statistical interim analysis

No statistical interim analysis is performed.

5.6 Time of outcome assessments

Data are collected during a single outpatient visit and are assessed within one week after study completion.

6. Statistical principles

6.1 Confidence interval and P values

A two-sided P value less than 0.05 is considered statistically significant. Two-side 95% confidence intervals (CIs) of between-group differences are reported for quantitative analysis.

6.2 Analysis population

All randomized participants were analyzed according to the intention-to-treat principle, in the groups to which they were originally assigned, regardless of protocol deviations.

6.3 Descriptive statistics

Continuous variables are summarized using median (interquartile range, IQR) or mean (SD), as appropriate.

Categorical variables are presented as frequencies and percentages.

7. Statistical methods

7.1 Demographic and baseline characteristics

Demographic data include sex, age, ocular conditions, and occupation status. Age is reported as median (IQR) and others as frequency (percentage). No statistical tests were performed to compare baseline characteristics, as randomization precludes systematic bias.

7.2 Inter-rater reliability

Inter-rater reliability is assessed using the intraclass correlation coefficient (ICC [2,1]) for continuous variables (MedHistory composite score) and weighted Cohen's κ for categorical variables (qualitative assessments of test recommendations). Scores from the two independent evaluators are averaged for efficacy analyses.

7.3 Primary efficacy analyses

The primary outcome is analyzed using multivariable linear regression, with study group (LLM-based agent vs ophthalmology residents) as the independent variable and age and sex included as covariates. Subscale scores (history-taking, communication, documentation quality) are analyzed as component measures.

Adjusted mean differences between groups are estimated and reported with 95% CIs and corresponding two-sided P values.

7.4 Secondary efficacy analyses

Secondary analyses 1: Patient satisfaction

Patient satisfaction is assessed using an 8-item questionnaire with responses recorded on a 5-point Likert scale (1–5). The first 7 items evaluate specific aspects of the pre-consultation experience, whereas the 8th item represents overall patient satisfaction. For each item, responses are summarized by frequency and percentage within each group, and median values with interquartile ranges (IQRs) are reported. Between-group comparisons for the satisfaction scores are performed using the Mann–Whitney U test, with median differences and corresponding two-sided P values reported.

Secondary analyses 2: Pre-consultation duration

Pre-consultation duration is compared between groups using the Mann–Whitney U test, due to anticipated non-normal distribution. Median differences and corresponding two-sided P values are reported.

Secondary analyses 3: Appropriateness of test recommendation

(1) Qualitative analyses

Qualitative scores are summarized as median (IQR) and as frequency and percentage distributions across score categories. Group differences are compared using the Mann–Whitney U test, with median differences and corresponding two-sided P values reported.

(2) Quantitative analyses

Quantitative performance is evaluated at the micro-level using recall, precision, and F1 score, computed across individual test-recommendation instances by aggregating true-positive, false-positive, and false-negative counts. 95% CIs and two-sided P values are estimated using bootstrap resampling (2000 iterations). Bootstrap resampling is stratified by study group, participants are resampled with replacement within each group, and group sizes are held constant across iterations.

7.5 Exploratory efficacy analyses

Exploratory analyses 1: Primary diagnosis performance

(1) Qualitative analyses

Qualitative scores are summarized as median (IQR) and as frequency and percentage distributions across score categories. Group differences are compared using the Mann–Whitney U test, with median differences and corresponding two-sided P values reported.

(2) Quantitative analyses

Quantitative performance at the micro-level is evaluated using recall, precision, and F1 score, computed across diagnostic instances by aggregating true-positive, false-positive, and false-negative counts. 95% CIs and two-sided P values are estimated using bootstrap resampling (2000 iterations). Bootstrap resampling is stratified by study group, participants are resampled with replacement within each group, and group sizes are held constant across iterations.

Macro-level analyses of primary diagnosis performance are conducted using linear mixed-effects models fitted at the participant level, with fixed effects for study group (LLM agent vs ophthalmology residents), evaluation stage (history-only vs signs-integration), and their interaction, and a participant-level random

intercept, adjusting for age and sex. Interaction terms are interpreted to assess differential changes in diagnostic performance across stages. 95% CIs and two-sided P values for fixed effects are estimated using parametric bootstrap resampling of model parameters (2000 iterations).

Exploratory analyses 2: Treatment plan formulation quality

Qualitative scores are summarized as median (IQR) and as frequency and percentage distributions across score categories. Group differences are compared using the Mann–Whitney U test, with median differences and corresponding two-sided P values reported.

Exploratory analyses 3: Projected test-related expenses and wait times

Projected test-related expenses and wait times are estimated using institutional data corresponding to the recommended tests. Waiting time is defined as the longest required duration among all recommended examinations. Alignment with the reference standard is assessed using absolute deviations, with lower values indicating closer alignment. Between-group comparisons of absolute deviations are performed using the Mann–Whitney U test. Effect sizes are summarized using Hodges–Lehmann estimates, defined as the median difference in absolute deviation between the resident and agent groups (resident minus agent), with corresponding 95% confidence intervals. cost and wait-time distributions are analyzed per category.

7.6 Handling of missing data

All primary, secondary, and exploratory analyses follow the intention-to-treat principle. Missing data are handled using multiple imputation.

7.7 Subgroup analysis

No subgroup analysis is performed.

7.8 Sensitivity analysis

No sensitivity analysis is performed.

7.9 Statistical software

Sample size is calculated in PASS 16 (NCSS, USA), other analyses are conducted using R software (version 4.0.3).

7.10 Deviations from the SAP

Any deviations from this SAP will be documented, justified, and reported in the final manuscript.

8. References

1. Rupert Bourne, GBD 2019 Blindness and Vision Impairment Collaborators, Vision Loss Expert Group of the Global Burden of Disease Study. Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study. *Lancet Glob Health*. 2021;9(2):e130-e143. doi:10.1016/S2214-109X(20)30425-3
2. Siegel H, Böhringer D, Wacker K, Josef Lopez Niedenhoff P, Mittelviefhaus H, Reinhard T. Duration of Consultations in an Outpatient Ophthalmology Unit. *Dtsch Arztebl Int*. 2023;120(27-28):481-482. doi:10.3238/arztebl.m2023.0037
3. Peterson MC, Holbrook JH, Von Hales D, Smith NL, Staker LV. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. *West J Med*. 1992;156(2):163-165.
4. Levine D. History taking is a complex skill. *BMJ*. Published online July 20, 2017;j3513. doi:10.1136/bmj.j3513
5. Zhakhina G, Tapinova K, Kanabekova P, Kainazarov T. Pre-consultation history taking systems and their impact on modern practices: Advantages and limitations. *J Clin Med Kazakhstan*. 2023;20(6):26-35. doi:10.23950/jcmk/13947
6. Young BK, Zhao PY. Large Language Models and the Shoreline of Ophthalmology. *JAMA Ophthalmol*. 2024;142(4):375-376. doi:10.1001/jamaophthalmol.2023.6937

7. Leypold T, Lingens LF, Beier JP, Boos AM. Integrating AI in Lipedema Management: Assessing the Efficacy of GPT-4 as a Consultation Assistant. *Life*. 2024;14(5):646. doi:10.3390/life14050646