Original Paper

A Multiagent Large Language Model Framework for Longitudinal Multimodal Tracking of Wet Age-Related Macular Degeneration: Retrospective Feasibility Study

Xuan Jiang[1], MS; Jinhua Tao[2], MD; Chunxia Yu[2], MD; Leilei Wang[2], MD; Xiaoling Fang[2], MD; Wenwen Xue[2], MD; Qin Qin[1], PhD

[1]School of Intelligent Manufacturing and Control Engineering, Shanghai Polytechnic University, Shanghai, China

[2]Shanghai Eye Disease Prevention & Treatment Center, Shanghai, China

Corresponding Author:
Jinhua Tao, MD
Shanghai Eye Disease Prevention & Treatment Center
380 Kangding Road
Shanghai, 200040
China
Phone: 86 13585656016
Email: jinhuatao@hotmail.com

Abstract

Background: Wet age-related macular degeneration (AMD) requires long-term, high-frequency management involving multimodal retinal imaging across repeated follow-up visits. Reconstructing disease trajectories from fragmented imaging reports, treatment histories, and clinical notes creates a substantial information integration burden for clinicians in routine chronic disease management.

Objective: This study aimed to evaluate the preliminary feasibility of a proof-of-concept, multiagent large language model (LLM) framework that integrates longitudinal multimodal imaging findings and treatment records into structured reports to support clinician review during wet AMD follow-up.

Methods: We developed a 3-layer multiagent LLM framework using GPT-4o and AutoGen. The framework consisted of modality-specific perception agents using a visual grounding prompting strategy, an integration agent, and a tracking report agent. To test technical feasibility, the framework was evaluated on a dense retrospective pilot dataset of 12 patients encompassing 69 long-term follow-up visits. AI-generated outputs were assessed against expert review at the visit level and through report-level quality scoring.

Results: In this exploratory cohort, the framework achieved a mean visit-level activity classification accuracy of 80.8% (SD 3.4%) across 4 independent runs (range 78.3%-85.5%), with a mean Cohen κ of 0.605 (SD 0.072; range 0.549-0.705). The framework generated structured longitudinal summaries of disease activity, multimodal findings, and treatment histories. The visual grounding prompting strategy reduced the nonimage reference rate from 13.8% (12/87) to 1.1% (1/87), supporting more consistent grounding of image-based outputs in visual evidence.

Conclusions: This proof-of-concept study supports the preliminary feasibility of using a multiagent LLM framework to organize longitudinal multimodal wet AMD follow-up data and produce structured summaries for clinician review. These reports are intended to support review of disease progression and treatment response associations during longitudinal follow-up. Future prospective multicenter studies with larger patient cohorts should evaluate whether these reports improve clinical workflow efficiency, reduce clinician workload, and support follow-up management decisions.

(JMIR Form Res 2026;10:e106685) doi: 10.2196/106685

KEYWORDS

age-related macular degeneration; large language model; optical coherence tomography; optical coherence tomography angiography; artificial intelligence; AI

XSL•FO
RenderX

Introduction

Wet age-related macular degeneration (AMD) is a chronic, sight-threatening retinal disease that requires long-term follow-up and repeated treatment decisions [1-3]. Anti–vascular endothelial growth factor (VEGF) therapy remains the mainstay of treatment [4], but its effectiveness depends on serial monitoring, repeated injections, individualized treatment intervals, and timely recognition of recurrent disease activity [5,6]. In routine care, disease activity is not judged from a single data point [7-9]. Clinicians must review serial fundus photographs, optical coherence tomography (OCT) images or report images, OCT angiography (OCTA) information when available, injection history, laterality, visual acuity, retinal thickness, and free-text visit notes [10-13]. Reconstructing this longitudinal trajectory across repeated visits creates a substantial information integration burden, particularly when the follow-up record is fragmented across imaging systems and clinical documentation formats [14].

This longitudinal review burden is different from single-image diagnosis. For wet AMD, the clinically relevant question is often whether activity has changed over time, whether structural findings are consistent across modalities, and whether anatomical response follows prior anti-VEGF treatment [3,15]. Missing modalities, inconsistent OCT or OCTA acquisition, variable report layouts, and mixed structured and unstructured clinical records make this task closer to clinical information synthesis than isolated image classification. A tool designed for this setting should therefore organize evidence for clinician review rather than replace specialist interpretation.

Recent multimodal large language models (LLMs) have shown potential for medical image interpretation, report generation, and multimodal clinical summarization [16-18]. However, many existing applications remain centered on single–time point classification or short-form diagnostic outputs [19,20]. Less attention has been paid to whether LLM-based systems can support longitudinal organization of chronic disease data [21], especially when the input consists of heterogeneous imaging reports, clinical records, and treatment histories collected during routine care [22-24]. This creates an opportunity for clinical informatics approaches that emphasize structured evidence extraction, missing data awareness, and clinician-facing summaries [25].

To address this gap, we developed a proof-of-concept multiagent LLM framework for longitudinal multimodal tracking of wet AMD follow-up data. The framework coordinates modality-specific agents for fundus photography and OCT and OCTA inputs with a longitudinal integration agent and structured reporting components. This study evaluated the preliminary feasibility of the framework in a small retrospective cohort, focusing on the framework's ability to organize fragmented multimodal evidence and generate structured longitudinal reports to support clinician review of disease progression and treatment response associations. The evaluation included visit-level activity classification and expert-rated report quality, with the aim of exploring the framework's potential to support longitudinal follow-up management.

Methods

Patients and Data Collection

This exploratory pilot study retrospectively collected multimodal retinal imaging data and corresponding electronic clinical records from patients diagnosed with wet AMD at the Shanghai Eye Disease Prevention and Treatment Center, Shanghai, China. To evaluate the framework's preliminary feasibility in chronic longitudinal tracking, the inclusion criteria were defined as follows: (1) a confirmed clinical diagnosis of wet AMD based on multimodal imaging and clinical examination, (2) receipt of at least one anti-VEGF injection, and (3) availability of longitudinal follow-up data with multimodal imaging or clinical records. Patients with insufficient follow-up information for chronological reconstruction, missing source images for expert review, severely degraded imaging quality that precluded specialist interpretation, or diagnoses that could not be reconciled with wet AMD–related macular neovascularization were excluded.

A dense pilot dataset of 12 unique patients encompassing 69 follow-up visits was constructed for this methodological evaluation. Imaging modalities captured at each visit included ultrawidefield fundus photography (Zeiss Clarus 700), OCT (Zeiss Cirrus or Heidelberg Spectralis, as available), and OCTA (Zeiss Cirrus, as available). Reflecting the pragmatic nature of real-world clinical workflows, the choice of OCT hardware was determined by device availability during routine care rather than by the study protocol.

The framework was evaluated in the same pragmatic setting in which the source data were generated. Not every visit contained all imaging modalities, and missing modalities were preserved rather than imputed. When a modality was unavailable or a converted report image was not interpretable, the corresponding agent returned a missing data or low confidence flag, and the integration agent generated the longitudinal summary from the remaining available evidence. This design was intended to reflect real-world follow-up documentation, where multimodal completeness varies across visits.

The baseline structured clinical records were extracted from the electronic medical database in Rich Text Format. Because the source clinical charts were documented in Chinese whereas the downstream multiagent framework operated in English, a standardized translation protocol was used: text fields were translated using a medical lexicon and subsequently cross-verified by 2 independent ophthalmic researchers to ensure semantic accuracy. To reduce potential information leakage, expert annotations used for evaluation were generated independently from the translated inputs provided to the framework. The framework inputs were limited to deidentified visit-level imaging materials and structured longitudinal fields needed for chronological reconstruction, including laterality, visit sequence, injection history, drug name, and basic clinical measurements; final expert activity labels and report-level scores were not included in any model input.

XSL•FO
RenderX

Ethical Considerations

This study was conducted in strict accordance with the tenets of the Declaration of Helsinki and was formally reviewed and approved by the institutional review board of the Shanghai Eye Disease Prevention and Treatment Center (approval EC-20230701-01). Informed consent was officially waived by the reviewing committee given the retrospective, noninterventional nature of the analysis.

To safeguard participant confidentiality, an extensive data deidentification pipeline was enforced before any computational analysis. All direct patient identifiers, including names, specific medical record numbers, and exact dates of birth, were permanently removed from the dataset. Images and clinical text were labeled only using randomized, deidentified study codes. Because retinal images and parsed clinical text were transmitted to a third-party commercial API, all transmitted content was reviewed to ensure that no identifiable patient information was included. This retrospective analysis was conducted under institutional review board approval with a waiver of informed consent; the computational workflow used only deidentified data for methodological evaluation.

No identifiable patient information was included in any API request payload. All longitudinal data transmissions used encrypted Transport Layer Security connections. Data handling complied with the enterprise data use terms in effect at the time of analysis, which specify secure end point processing and zero retention, meaning that transmitted inputs are not used by the commercial API service provider for model training, continuous fine-tuning, or service improvement.

Data Preprocessing Pipeline

A structured data preprocessing pipeline was developed to prepare heterogeneous imaging and clinical inputs for downstream multiagent evaluation. To process raw imaging documentation, PDF OCT and OCTA clinical reports were programmatically converted to high-resolution JPEG image files using PyMuPDF (Artifex). In this manuscript, “OCT reports” and “OCTA reports” refer to converted report images containing retinal scan panels and machine-generated quantitative displays rather than free-text clinical interpretations. To reduce identity leakage during external API transmission, an automated spatial cropping protocol was applied to remove headers, footers, and metadata-containing regions before analysis.

Concurrently, text-based raw clinical records, which were originally documented in Chinese, were parsed using regular expressions tailored to 5 distinct longitudinal injection record formats and documentation styles across visits. This pipeline extracted raw textual metrics, including best-corrected visual acuity, intraocular pressure, injection history, drug names, injection dates, and follow-up recommendations, which were subsequently translated and mapped into structured visit-level records.

Framework Architecture

The proposed proof-of-concept framework follows a modular, 3-layer multiagent architecture designed to emulate a collaborative clinical information review workflow. It comprises a perception layer, a longitudinal integration layer, and a structured report layer. The overall methodological workflow and agent interactions are illustrated in Figure 1.

In the perception layer, 3 modality-specific vision language agents independently analyze fundus photographs, OCT report images, and OCTA report images at each visit time point. Each agent operates within an isolated conversational session with zero cross-modality context sharing, minimizing confirmation bias during initial image interpretation. The fundus agent performs automated image quality control and artifact detection, including assessment of specular reflection and uneven illumination, before evaluating macular findings related to disease activity. The OCT and OCTA agents extract structured imaging features from the available report images while explicitly flagging missing, low-quality, or noninterpretable inputs.

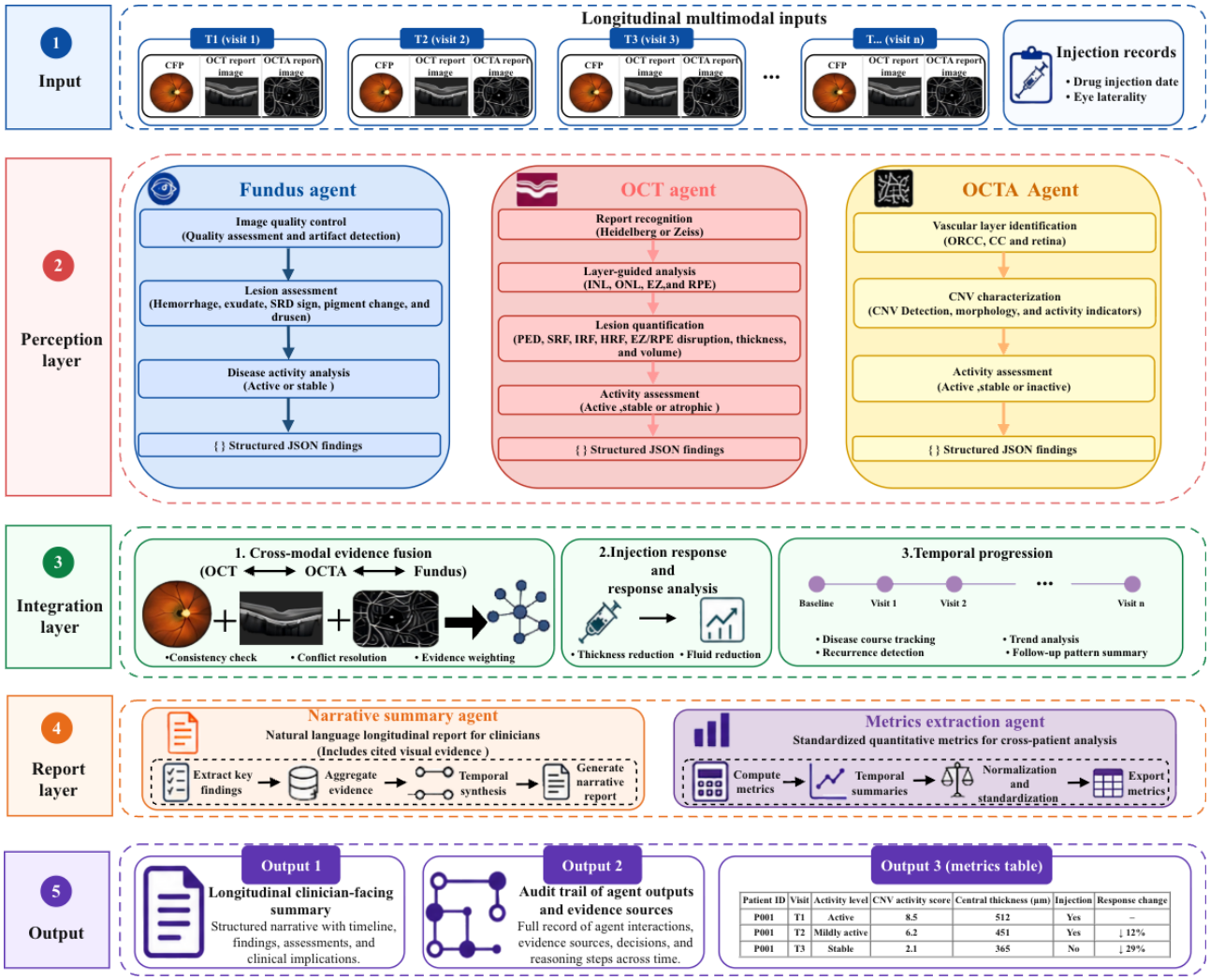
In the integration layer, a longitudinal agent synthesizes visit-level findings across all available follow-up time points for each patient. The agent performs cross-modal evidence fusion; reconciles potentially conflicting findings across fundus photography, OCT, and OCTA through consistency checking and evidence weighting; and links injection events to subsequent structural changes with explicit awareness of injection dates, drug names, laterality, and follow-up intervals.

In the report layer, a report agent generates a structured longitudinal clinical summary in natural language, including per-visit activity overviews, detailed multimodal interpretations, longitudinal trend summaries, treatment response analysis, and data quality limitations. In parallel, a metrics agent extracts standardized visit-level structured fields for downstream cross-patient comparison and summary analysis.

GPT-4o served as the underlying model for all agents and was accessed through an OpenAI-compatible commercial API service [26]. The multiagent framework was implemented using the Microsoft AutoGen library [27]. During development and manuscript preparation, the framework was run repeatedly to refine prompt templates, structured output schemas, and failure-handling rules. For the formal feasibility evaluation reported in this study, the prompt templates, structured output schemas, and API configuration were locked before case-level evaluation and held constant across all independent runs. The reported results were based exclusively on outputs generated under this locked configuration from deidentified multimodal inputs. Within each independent evaluation run, each available modality at each visit was processed once. A checkpoint-based resumption mechanism allowed interrupted analyses to continue from the last completed visit without reprocessing completed visits. API keys, account credentials, billing information, and private end point credentials were not shared. The model configuration, prompt templates, structured JSON output schema, and evaluation rubric are provided in Multimedia Appendix 1. The framework was designed as a clinician-facing information organization tool rather than an autonomous diagnostic system; final clinical interpretation and management decisions remained the responsibility of ophthalmologists.

XSL•FO
RenderX

Figure 1. Multiagent large language model workflow for longitudinal multimodal wet age-related macular degeneration follow-up summarization. The framework integrates longitudinal fundus photography, optical coherence tomography (OCT), OCT angiography (OCTA), and injection records. Modality-specific agents extract structured findings, which are integrated across modalities and visits to generate clinician-facing longitudinal summaries, audit trails, and quantitative metrics. CC: choriocapillaris; CFP: color fundus photography; CNV: choroidal neovascularization; EZ: ellipsoid zone; HRF: hyperreflective foci; INL: inner nuclear layer; IRF: intraretinal fluid; ONL: outer nuclear layer; ORCC: outer retinal–choriocapillaris complex; PED: pigment epithelial detachment; RPE: retinal pigment epithelium; SRD: serous retinal detachment; SRF: subretinal fluid.



Visual Grounding Prompting Strategy

A visual grounding prompting strategy was designed for each modality-specific agent to guide structured, evidence-based image interpretation, informed by prior work on stepwise prompting and medical reasoning prompts [28,29]. Rather than asking the model to produce a direct diagnosis, each prompt instructed the agent to follow an explicit sequence of observable visual checks before reaching a structured conclusion. Figure 2 illustrates this workflow using the OCT agent as a representative example.

For the fundus agent, the interpretation sequence began with image quality and artifact assessment, including specular reflection, uneven illumination, and media opacity. The agent then localized the macular region and systematically examined major activity-related signs, including hemorrhage, hard or soft exudation, serous detachment signs, pigmentary change, edema-related elevation, choroidal neovascularization (CNV) membrane appearance, and geographic atrophy, distinguishing active from chronic findings where applicable.

For the OCT agent, the interpretation sequence followed six explicit steps: (1) report type identification (Heidelberg B-scan, Heidelberg thickness map, or Zeiss OCT) as these formats differ in how structural and quantitative information is presented; (2) image quality assessment, with analysis halted if the quality score fell below 5/10; (3) layer-guided structural analysis of the fovea, inner nuclear layer, outer nuclear layer, ellipsoid zone, and retinal pigment epithelium; (4) priority-based lesion detection following the order of pigment epithelial detachment (PED), subretinal fluid (SRF), intraretinal fluid, hyperreflective foci, and ellipsoid zone disruption; (5) extraction of central retinal thickness when available; and (6) structured activity classification.

For the OCTA agent, the prompt first required identification of the displayed vascular layer based on the highlighted panel within the report. The agent then interpreted flow information according to the identified layer, with particular emphasis on the avascular zone for CNV detection. Activity-related features, including arborizing pattern, feeder vessels, and peripheral

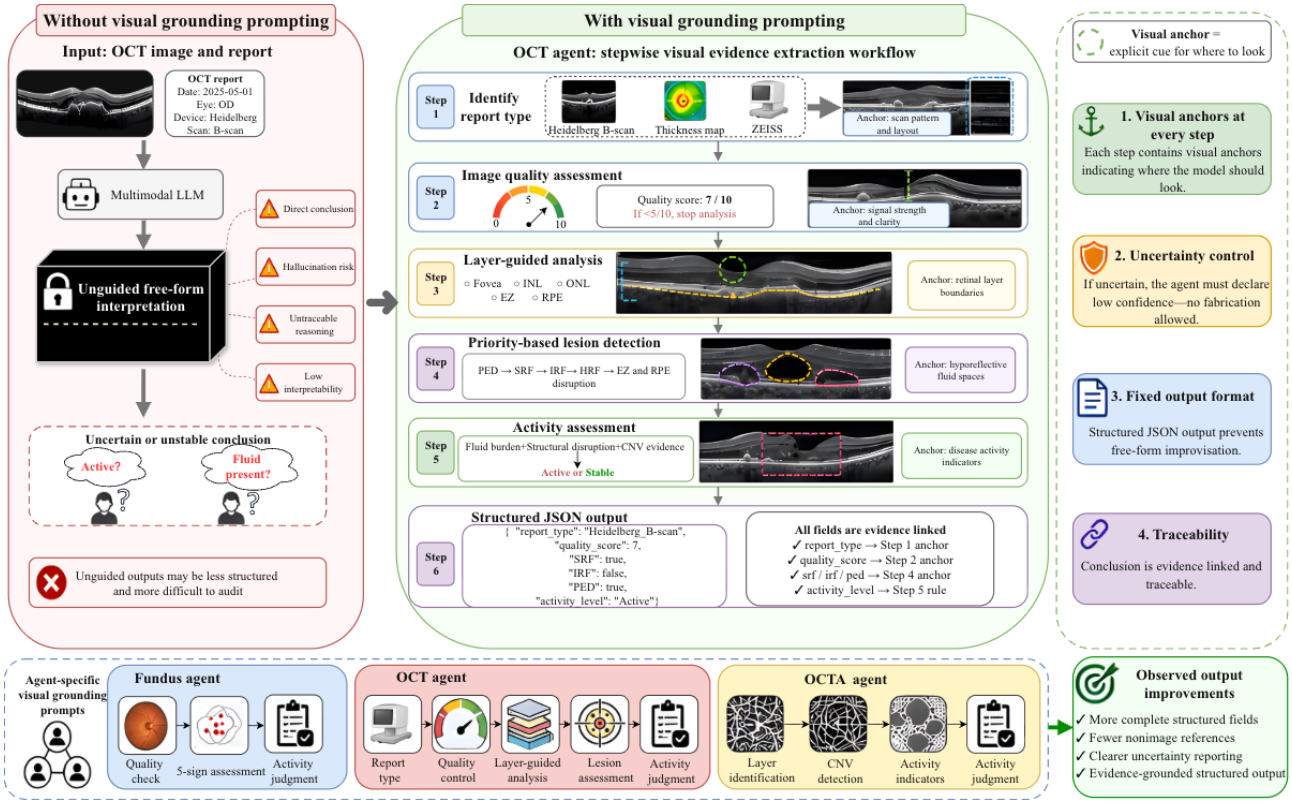
XSL•FO
RenderX

anastomosis rings, were assessed, while projection artifacts were explicitly distinguished from true pathological flow.

All agents were instructed to output results in JSON format and include an image quality score and confidence level for each assessment. Historical summaries were selectively injected to provide longitudinal context without exceeding token limits, but final judgments for each modality remained grounded primarily in the current image.

For reproducibility, the operational workflow was defined by modality-specific prompt templates, a structured JSON output schema, fixed preprocessing steps, and predefined handling rules for missing or failed modality outputs. The complete prompts and schema are not included in the main text because of length constraints but are suitable for inclusion as supplementary material or a deidentified repository accompanying submission.

Figure 2. Visual grounding prompting workflow for structured optical coherence tomography (OCT) interpretation in the multiagent framework. The left panel shows the limitations of unguided prompting, including direct conclusions, hallucination risk, low traceability, and limited interpretability. The right panel illustrates the stepwise OCT-agent workflow, including report type identification, image quality assessment, layer-guided analysis, lesion detection, activity assessment, and structured JSON generation. Visual anchors, uncertainty control, fixed output formats, and evidence-linked traceability support more complete and interpretable outputs across the fundus, OCT, and OCT angiography (OCTA) agents. CNV: choroidal neovascularization; EZ: ellipsoid zone; HRF: hyperreflective foci; INL: inner nuclear layer; IRF: intraretinal fluid; LLM: large language model; OD: oculus dexter (right eye); ONL: outer nuclear layer; PED: pigment epithelial detachment; RPE: retinal pigment epithelium; SRF: subretinal fluid.



Expert Evaluation

To evaluate the framework as a clinician-facing information organization tool, we performed a 2-level expert assessment.

First, time point–level disease activity summaries generated by the framework were compared with expert annotations across all follow-up visits from 12 patients. The primary evaluation unit was the follow-up time point for the study eye under longitudinal review. Disease activity was classified as active when the expert identified imaging evidence consistent with active wet AMD, such as SRF, intraretinal fluid, new or enlarged PED, hemorrhage, active macular neovascular flow on OCTA when available, or clinically meaningful worsening of retinal thickness in the context of treatment history. Stable disease was defined as the absence of new activity-related findings or the presence of chronic inactive structural changes without evidence of exudative worsening. This analysis was used to assess whether the framework could produce visit-level activity summaries that were broadly consistent with specialist review.

Second, report-level quality evaluation was conducted on 4 representative information-rich patients (P001, P006, P008, and P012) with relatively complete longitudinal multimodal follow-up data. These cases were selected before scoring to test whether the framework could synthesize longer and more information-dense follow-up trajectories; therefore, this evaluation should be interpreted as a qualitative report quality assessment rather than an unbiased estimate of performance across all patients. A retinal specialist reviewed the unedited AI-generated longitudinal reports and rated them on four dimensions using a 5-point Likert scale: (1) accuracy of activity assessment, (2) multimodal evidence integration, (3) accuracy of longitudinal disease course and treatment response summarization, and (4) perceived value for long-term follow-up management support.

XSL•FO
RenderX

Statistical Analysis

Descriptive statistics were used to summarize dataset characteristics, modality availability, visit-level activity classification performance, report-level expert scores, and ablation metrics. Visit-level activity assessment was evaluated using accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and the Cohen κ. The 95% CIs for binomial proportions were calculated using the Wilson score method, and the CI for the Cohen κ was calculated using an asymptotic SE approach [30]. For the repeated-run analysis, performance variability was summarized using the mean, SD, and range across the 4 independent executions. Because this was a small retrospective feasibility study with repeated visits nested within patients, the analyses were interpreted descriptively, and no formal hypothesis testing or patient-level clustering adjustment was performed.

Results

Dataset Characteristics

The study dataset comprised 12 patients with wet AMD recruited from the Shanghai Eye Disease Prevention and Treatment Center. All patients underwent longitudinal multimodal retinal imaging as part of routine clinical follow-up, including fundus photography, OCT, and OCTA. Imaging modality availability varied across visits because of routine clinical scheduling, with OCTA being less consistently acquired than fundus photography and OCT. OCT examinations were obtained using Zeiss Cirrus or Heidelberg Spectralis devices according to routine clinical availability, and longitudinal comparisons therefore reflected pragmatic real-world data rather than a standardized imaging protocol. Dataset characteristics are summarized in Table 1.

Table 1. Characteristics of the study dataset.

Variables	Values
Patients, n	12
Follow-up visits, n	69
Follow-up visits per patient, median (IQR; range)	5.5 (5.00-6.25;3-9)
Follow-up duration (mo), median (IQR; range)	10.2 (6.83-14.26; 4.9-16.1)
Visits with fundus photography available (n=69), n (%)	66 (95.7)
Visits with OCT[a] available (n=69), n (%)	57 (82.6)
Visits with OCTA[b] available (n=69), n (%)	31 (44.9)
Visits with all 3 modalities available (n=69), n (%)	29 (42.0)
Visits with incomplete modality data (n=69), n (%)	40 (58.0)

[a]OCT: optical coherence tomography.

[b]OCTA: OCT angiography.

Visit-Level Activity Assessment Validation

To evaluate the framework's disease activity classification at the visit level, we compared its outputs against independent annotations provided by a retinal specialist with more than 10 years of clinical experience in macular degeneration. The expert reviewed the original imaging data for each visit without access to the framework's reports, classifying each time point as active or stable based on the predefined clinical criteria described in the Methods section. The same expert performed report-level scoring after the visit-level reference labels had been finalized; this potential source of evaluation bias is acknowledged as a limitation.

In the primary evaluation (run 1), across all 69 time points from 12 patients, the framework achieved an overall accuracy of 81.2% (56/69; 95% CI 70.4%-88.6%) and a Cohen κ of 0.609 (95% CI 0.418-0.801), indicating agreement beyond chance with the expert reference. Sensitivity was 89.7% (95% CI 76.4%-95.9%), and specificity was 70.0% (95% CI 52.1%-83.3%), with a positive predictive value of 79.5% (95% CI 65.5%-88.8%) and a negative predictive value of 84.0% (95% CI 65.3%-93.6%). These results support preliminary agreement with expert review rather than definitive evidence of clinical effectiveness (Table 2).

XSL•FO
RenderX

Table 2. Visit-level activity assessment performance in the primary evaluation (run 1). Performance was evaluated against retinal specialist reference labels across 69 visits from 12 patients.

Metrics	Values
Total time points, n	69
Overall accuracy, n/N (%; 95% CI)	56/69 (81.2; 70.4-88.6)
Cohen κ (95% CI)	0.609 (0.418-0.801)
Sensitivity (%; 95% CI)	89.7 (76.4-95.9)
Specificity (%; 95% CI)	70.0 (52.1-83.3)
Positive predictive value (%; 95% CI)	79.5 (65.5-88.8)
Negative predictive value (%; 95% CI)	84.0 (65.3-93.6)
True positives, n	35
True negatives, n	21
False positives, n	9
False negatives, n	4

Report-Level Expert Evaluation

To assess the quality of the generated longitudinal reports at the patient level, a structured 4-dimension evaluation form was developed based on a 5-point Likert scale (1="strongly disagree"; 5="strongly agree"), covering activity assessment accuracy, multimodal evidence integration, longitudinal disease course and treatment response summarization, and long-term follow-up management support value. The Likert scale was selected because of its established use in clinical tool evaluation and its ability to capture expert judgment across multiple dimensions [31].

Four representative patients (P001, P006, P008, and P012) with relatively complete longitudinal multimodal data were reviewed by a retinal specialist using this form. The mean scores were 4.00 (SD 0.00) for activity assessment accuracy, 4.00 (SD 0.00) for multimodal evidence integration, 3.75 (SD 0.50) for longitudinal disease course and treatment response summarization, and 3.25 (SD 0.50) for long-term follow-up management support. The overall mean score was 3.75 (SD 0.20) out of 5, as detailed in Table 3.

Table 3. Expert evaluation scores across 4 representative patients. Scores were assigned on a 5-point Likert scale.

Dimension	P001	P006	P008	P012	Scores, mean (SD)[a]
Activity assessment accuracy	4	4	4	4	4.00 (0.00)
Multimodal evidence integration	4	4	4	4	4.00 (0.00)
Longitudinal disease course summarization	4	4	3	4	3.75 (0.50)
Follow-up management support	3	3	3	4	3.25 (0.50)
Values, mean (SD)	3.75	3.75	3.5	4.00	3.75 (0.20)

[a]All SDs were calculated using the sample SD formula (n−1) across the 4 patients. The overall SD was calculated from the 4 patient-level mean scores.

Additional Experiments

Effect of Visual Grounding Prompting

To evaluate the contribution of the visual grounding prompting strategy, we conducted an ablation study comparing the full framework against a variant without visual grounding prompting in which each agent received only a minimal instruction to analyze the image and return results in JSON format. Both conditions were evaluated across 87 modality-level records from 6 patients (P001-P006), covering fundus and OCT and OCTA agents. This ablation was designed to assess output structure and evidence grounding rather than downstream clinical accuracy.

As shown in Table 4, the visual grounding prompting condition achieved higher performance across all structural output metrics. JSON completeness was defined as the proportion of outputs that could be parsed into the expected schema. Key field missingness rate captured absent required fields. Confidence declaration rate and image quality score inclusion rate measured whether the output included the requested uncertainty and image quality fields. Activity enumeration compliance measured whether activity labels used predefined categories. Nonimage reference rate was judged by reviewing confidence justifications and identifying outputs that cited nonvisual contextual information rather than current image evidence. This rate decreased from 13.8% (12/87) without visual grounding prompting to 1.1% (1/87) with visual grounding prompting.

XSL•FO
RenderX

Table 4. Structural output and evidence-grounding metrics for visual grounding prompting ablation. The table compares the full visual grounding prompting condition with the condition without visual grounding across 87 modality-level records from 6 patients (n=87)[a].

Metrics	With visual grounding prompting	Without visual grounding prompting
Required field completeness (%), mean (SD)[b]	91.6 (25.93)	85.5 (34.60)
Per-record key field missingness rate (%), mean (SD)	10.1 (29.89)	13.8 (34.68)
Reports with confidence declaration, n (%)	79 (90.8)	75 (86.2)
Reports with image quality score, n (%)	80 (92.0)	75 (86.2)
Reports with valid activity enumeration, n (%)	77 (88.5)	73 (83.9)
Reports with nonimage reference, n (%)	1 (1.1)	12 (13.8)

[a]A total of 87 modality- and eye-specific outputs were evaluated. Required field completeness and key field missingness rate were calculated within each output according to its modality-specific schema and are reported as mean per-record percentages; they are not proportions of records out of 87.

[b]All SDs were calculated across the 87 records using the sample SD formula (n−1). The recalculated means are consistent with the originally reported values.

Comparison With a Rule-Based Template Baseline

To examine the contribution beyond deterministic record summarization, we implemented a rule-based extraction and fixed-template baseline for P001, P006, P008, and P012. This baseline organized available clinical records into chronological summaries of visual acuity, documented examination findings, injection history, and follow-up information. Date intervals and numerical changes were calculated when supported by the source data. Imaging findings were reproduced from existing clinical descriptions rather than independently extracted from images.

The rule-based baseline demonstrated that basic timeline construction and aggregation of documented information could be achieved without an LLM. The multiagent framework extended this scope by incorporating modality-specific image interpretation, explicit comparison of evidence across modalities, and longitudinal synthesis of imaging findings alongside treatment records. Its reports included descriptions of image quality, lesion features, quantitative measurements where available, and modality-specific limitations, providing additional information for clinician review beyond the recorded clinical text. Thus, the intended added value of the multiagent approach lies not simply in generating a narrative summary but also in organizing image-derived evidence and its relationship to the longitudinal clinical record. The reliability of these additional interpretations remains dependent on their agreement with the source images and clinical assessment.

Comparison With a Single-Agent LLM Baseline

To examine the contribution of the multiagent organization beyond the use of an LLM alone, we designed a single-agent baseline comparison involving the same 4 patients included in the original report-level evaluation (P001, P006, P008, and P012). The baseline was configured to use GPT-4o to perform image interpretation, cross-modal comparison, temporal integration, and report generation within a single response, without modality-specific agents or structured interagent handoffs. The prompt requested modality-specific findings, evidence sources, image quality limitations, longitudinal changes, and treatment response associations, with separate consideration of each eye.

As shown in Table 5, the multiagent framework received higher mean expert ratings than the single-agent baseline across all 4 dimensions. The largest numerical differences were observed in activity assessment accuracy (4.00 vs 2.75) and multimodal evidence integration (4.00 vs 3.00), followed by longitudinal disease course summarization (3.75 vs 3.00). The difference in follow-up management support was smaller (3.25 vs 3.00). These descriptive findings suggest that, in the 4 evaluated cases, the perceived benefit of the multiagent framework was more pronounced in activity assessment and multimodal evidence integration than in follow-up management support. However, the small sample, single evaluator, and separate scoring rounds limit conclusions about comparative superiority.

XSL•FO
RenderX

Table 5. Patient-matched expert evaluation of multiagent and single-agent longitudinal reports across 4 patients[a].

Dimension and method	P001	P006	P008	P012	Mean (SD)[b]
Activity assessment accuracy					
Multiagent	4	4	4	4	4.00 (0.00)
Single agent	3	2	3	3	2.75 (0.50)
Multimodal evidence integration					
Multiagent	4	4	4	4	4.00 (0.00)
Single agent	3	3	3	3	3.00 (0.00)
Longitudinal disease course summarization					
Multiagent	4	4	3	4	3.75 (0.50)
Single agent	3	2	3	4	3.00 (0.82)
Follow-up management support					
Multiagent	3	3	3	4	3.25 (0.50)
Single agent	3	3	3	3	3.00 (0.00)

[a]Scores range from 1 to 5, with higher scores indicating more favorable expert assessments according to the dimension-specific criteria. Mean values represent the arithmetic mean across the 4 patients for each method and dimension. Multiagent scores were obtained in the original evaluation, whereas single-agent scores were obtained in a subsequent evaluation by the same retinal specialist. Comparisons are descriptive and exploratory.

[b]All SDs were calculated using the sample SD formula (n−1) across the 4 patients.

Run-to-Run Stability and Reproducibility

To assess the run-to-run variability and stability of the proposed framework, the complete multiagent pipeline was independently executed 4 times on the same dataset comprising 12 patients and 69 visits.

The input data, prompts, model configuration, activity classification criteria, and evaluation procedure were held constant across all 4 runs. Each execution independently generated new modality-specific findings and integrated outputs without reusing responses from previous runs. Predictions from each run were evaluated against the same retinal specialist reference labels using an identical eye-to-visit aggregation rule. Run-to-run variability was summarized using the mean, SD, and range of the accuracy, sensitivity, specificity, and Cohen κ across the 4 independent runs. The performance of each individual run and the corresponding summary statistics are shown in Table 6.

As shown in Table 6, performance varied across the 4 independent runs, with accuracy ranging from 78.3% to 85.5% (mean 80.8%, SD 3.4%). Sensitivity showed relatively limited variation across runs (mean 86.5%, SD 3.2%), whereas greater variability was observed in specificity (mean 73.3%, SD 7.2%). The Cohen κ ranged from 0.549 to 0.705. Despite measurable run-to-run variability, the framework maintained broadly comparable overall classification performance across the 4 independent executions.

Table 6. Run-to-run variability in activity classification performance across 4 independent executions of the multiagent framework[a].

Metrics	Run 1	Run 2	Run 3	Run 4	Mean (SD)	Range
Accuracy (%)	81.2	85.5	78.3	78.3	80.8 (3.4)	78.3-85.5
Sensitivity (%)	89.7	87.2	87.2	82.1	86.5 (3.2)	82.1-89.7
Specificity (%)	70.0	83.3	66.7	73.3	73.3 (7.2)	66.7-83.3
Cohen κ	0.609	0.705	0.549	0.556	0.605 (0.072)	0.549-0.705

[a]Mean values, SDs, and ranges summarize performance variability across the 4 runs. Sensitivity and specificity represent the correct classification of active and stable visits, respectively.

Illustrative Cases

Patient P012 illustrates the framework's ability to identify multimodal discordance for clinician review. At visit 1, fundus photography suggested stable disease in the right eye, while OCT revealed a fibrovascular PED with a height of approximately 250 µm, SRF, and a central retinal thickness of 360 µm. The framework classified the visit as active based on OCT evidence despite the stable fundus appearance. A representative excerpt from the AI-generated report is shown in Textbox 1.

XSL•FO
RenderX

Textbox 1. Representative excerpt from the AI-generated longitudinal report for patient P012 at visit 1 (the baseline visit). The report integrates fundus photography, optical coherence tomography (OCT) and OCT angiography (OCTA) findings. Although fundus photography suggested stable disease, OCT identified a fibrovascular pigment epithelial detachment (PED) and subretinal fluid (SRF) in the right eye. Based on the multimodal evidence, the framework classified the right eye as active and the left eye as stable.

Visit 1 (Baseline, no prior in

with slight irregularity, consistent with chronic change

- CNV membrane: Not identified

→ OD fundus assessment: Stable. Findings limited to chronic hard exudates and pigmentary changes; no fresh hemorrhage or active leakage.

· OS:

- Hemorrhage: None

- Exudate: Diffuse chronic hard exudates in and around the macula, consistent with lipid deposits; no active features

- Foveal reflex: Present

- Pigmentary changes: Mild focal depigmentation with slight irregularity

- CNV membrane: Not identified

→ OS fundus assessment: Stable. No features consistent with active leakage.

(OCT) (Confidence: High; Report type: Zeiss OCT)

· Image quality: 9/10, clear image with well-defined retinal layers.

· OD:

- PED: Fibrovascular PED clearly identified beneath the RPE, height approximately 250 μm, indicative of active lesion

- SRF: Small amount of subretinal fluid present superior to the PED

- IRF: Not identified

- HRF: Scattered hyperreflective foci

- Central retinal thickness: 360 μm (markedly increased)

- Ellipsoid zone (EZ): Focal disruption, indicating photoreceptor damage

- RPE layer: Continuity compromised by pronounced elevation

→ OD OCT assessment: Active. Primary evidence: fibrovascular PED with SRF, consistent with active CNV.

· OS:

- PED: Not identified

- SRF: Not identified

- IRF: Not identified

- Central retinal thickness: 260 μm

- EZ: No significant abnormality

- RPE layer: No significant abnormality

→ OS OCT assessment: Stable. Overall macular structure normal; no exudative signs.

(OCTA) (Confidence: High; Layer analyzed: Retinal layer)

· OD: Image quality 9/10. No abnormal flow signal detected; no CNV identified. Note: current layer is the retinal layer, which has limited sensitivity for direct visualization of deeper CNV.

· OS: Image quality 8/10. No abnormal flow signal detected; no CNV identified. Same layer limitation applies.

→ OCTA assessment: No definitive CNV flow signal in either eye. Interpretation limited by layer selection; OCT findings must be considered in conjunction.

(Multimodal Consistency)

Fundus photography of OD suggested chronic stable changes; however, OCT clearly demonstrated fibrovascular PED and SRF, indicating that structural imaging is more sensitive than fundus appearance for detecting active disease. OD was therefore classified as active based on OCT evidence. OS findings were consistent across all three modalities, supporting a stable classification.

(Injection Status)

Baseline visit. No prior anti-VEGF injection. Given active disease in OD, treatment was initiated following this visit.

XSL•FO
RenderX

Beyond single-visit classification, the framework's longitudinal value lies in its ability to synthesize trends and treatment response associations across visits, a capability not accessible from single–time point analysis alone. For patient P012, the report layer automatically tracked right-eye central retinal thickness across 5 visits (360, 258, 224, 237, and 235 µm, respectively) and related these changes to the timing of 4 anti-VEGF injections, identifying the most pronounced anatomical improvement after early treatment (Table 7).

Table 7. Structured per-visit summary from the AI-generated longitudinal report for patient P012 (right eye). The table summarizes findings across 5 visits, including time since the last injection, fundus findings, optical coherence tomography (OCT)–based central retinal thickness (CRT) and lesion status, OCT angiography (OCTA) findings, and the resulting composite activity classification.

Visit	Days since last injection	Fundus (OD[a])	OCT (CRT; µm)	OCTA	Composite activity
T1	—[b] (baseline)	No active features	Active (PED[c]+SRF[d]); 360	No CNV[e] detected	Active
T2	20	Stable	Stable; 258 (−102)[f]	Parsing failed	Stable
T3	20	Stable	Stable; 224 (−34)[f]	No CNV signal	Stable
T4	58	Stable	Stable; 237 (+13)[f]	Parsing failed	Stable
T5	35	Stable	Stable; 235 (−2)[f]	No CNV signal	Stable

[a]OD: oculus dexter (right eye).

[b]Not applicable.

[c]PED: pigment epithelial detachment.

[d]SRF: subretinal fluid.

[e]CNV: choroidal neovascularization.

[f]Values in parentheses indicate changes in CRT (m) relative to the immediately preceding visit. Negative values indicate a decrease in CRT, whereas positive values indicate an increase.

Discussion

Principal Findings

This study presented a multiagent LLM framework for longitudinal multimodal tracking of wet AMD, with the goal of supporting long-term follow-up management rather than automating single-visit diagnosis. By integrating fundus photography, OCT, and OCTA findings across repeated visits, the framework generated structured reports summarizing disease trajectory and treatment response. The main contribution of this study is not the automation of wet AMD diagnosis but the demonstration of a feasible informatics workflow for organizing longitudinal multimodal evidence for clinician review.

Across 4 independent runs, the framework achieved a mean visit-level activity classification accuracy of 80.8% (SD 3.4%; range 78.3%-85.5%) and a mean Cohen κ of 0.605 (SD 0.072; range 0.549-0.705). Sensitivity remained relatively consistent across runs (mean 86.5%, SD 3.2%), whereas greater variability was observed in specificity (mean 73.3%, SD 7.2%). These findings indicate measurable run-to-run variability while showing broadly comparable aggregate classification performance across repeated executions. Given the small single-center pilot cohort, these results should be interpreted as preliminary evidence of feasibility rather than definitive clinical performance.

The ablation study suggested that visual grounding prompting improved the structural reliability of modality-specific outputs. The longitudinal integration capability enabled the framework to capture disease trajectories and injection-response associations that are structurally inaccessible from single-visit analysis alone. The visual grounding prompting strategy improved evidence grounding by reducing the nonimage reference rate from 13.8% (12/87) to 1.1% (1/87); its effect on downstream diagnostic accuracy was not directly tested here and warrants dedicated evaluation.

Comparison With Prior Work

Previous studies have primarily evaluated multimodal LLMs for single-image or single-visit retinal disease classification and report generation. In contrast, the present study focused on longitudinal organization of multimodal follow-up information, treatment history, and disease activity summaries for clinician review. This distinction is important because chronic wet AMD management requires repeated reconstruction of disease trajectories rather than isolated image interpretation. The findings therefore complement prior work by exploring an informatics workflow for longitudinal evidence synthesis while remaining preliminary because prospective workflow efficiency and clinical decision impact were not directly measured.

Limitations

Several limitations should be acknowledged. The dataset comprised 12 patients from a single center, limiting generalizability. Imaging was acquired during routine care rather than under a standardized prospective protocol; OCT devices varied across Zeiss Cirrus and Heidelberg Spectralis systems, and the available OCT report images may not capture all B-scan locations. Therefore, parafoveal fluid or lesions outside the displayed scan panels could have been missed. OCTA availability was inconsistent across visits, and the displayed or analyzed layer was not always the avascular zone, reducing CNV assessment reliability at certain time points. A second limitation is the run-to-run variability inherent in the LLM-based framework. Although the complete pipeline was independently

XSL•FO
RenderX

executed 4 times under identical experimental conditions, measurable variation remained across runs, particularly in specificity. Further validation with larger datasets and additional repeated executions is needed to better characterize the reproducibility of the framework.

In addition, a single retinal specialist performed both the visit-level reference annotations and the subsequent report-level quality scoring. Although the specialist was blinded to the framework-generated reports during visit-level reference annotation, the single-rater design precluded formal assessment of interrater reliability. Moreover, report-level scoring necessarily involved access to the generated reports and may have introduced evaluator bias. Future validation should involve multiple independent retinal specialists, assess interrater agreement, and use consensus or adjudicated reference labels where appropriate. The ablation analysis assessed output structure and evidence grounding but did not directly test whether visual grounding prompting improved clinical activity classification. Because this was a small retrospective feasibility study, the findings should be interpreted as evidence of methodological feasibility rather than clinical effectiveness.

Although the framework is intended to support longitudinal follow-up management, this study did not directly measure clinician review time, workload, or changes in management decisions. The expert ratings assessed report quality and perceived follow-up support rather than actual burden reduction or improvements in patient outcomes. Future studies should include larger and more diverse patient cohorts and multiple clinicians comparing review with and without AI-generated reports. Outcomes should include review time, workload, and the accuracy and completeness of identifying clinically relevant information.

Conclusions

This retrospective feasibility study provides preliminary evidence that the framework can generate longitudinal multimodal reports to support clinician review in wet AMD follow-up. Larger studies with direct workflow evaluation are needed to determine whether these reports reduce clinician burden or improve follow-up management.

Funding

The authors declared no financial support was received for this work.

Data Availability

The original patient imaging data and clinical records are not publicly available because they are subject to institutional privacy, ethical, and data sharing restrictions. Requests for access to deidentified data may be directed to the corresponding author and will be considered on a case-by-case basis subject to institutional review board approval and institutional data sharing policies. Nonidentifiable study materials, including representative prompt templates, structured output schemas, model configuration, evaluation rubrics, and aggregate performance metrics, are available in the supplementary materials.

Generative AI Use

The authors declare that generative AI was used in the creation of this manuscript. Generative AI (ChatGPT) was used for proofreading, editing, or language polishing; translation. The authors verified all outputs of the tool(s) used in the preparation of this manuscript and take full responsibility for the final content of the manuscript. GPT-4o was also used as the underlying large language model in the proposed framework, as described in the Methods section.

Authors' Contributions

Conceptualization: XJ, JT
Data curation: XJ, CY, LW, XF, WX
Formal analysis: XJ, QQ
Investigation: XJ, CY, LW, XF, WX
Methodology: XJ, QQ, JT
Project administration: JT
Software: XJ
Supervision: JT, QQ
Validation: CY, LW, XF, WX, JT
Visualization: XJ
Writing—original draft: XJ
Writing—review and editing: XJ, JT, QQ

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1

Prompt templates, structured output schema, model configuration, and evaluation rubric.
[DOCX File , 32 KB-Multimedia Appendix 1]

References

1. Wong WL, Su X, Li X, Cheung CM, Klein R, Cheng CY, et al. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. Lancet Glob Health. Feb 2014;2(2):e106-e116. [FREE Full text] [doi: 10.1016/S2214-109X(13)70145-1] [Medline: 25104651]
2. Stoica AM, Jurja S, Dervis N. Anti-VEGF therapy - a new hope in AMD treatment. Rom J Ophthalmol. 2025;69(3):322-327. [doi: 10.22336/rjo.2025.52] [Medline: 41189785]
3. Abedi F, Wickremasinghe S, Islam AF, Inglis KM, Guymer RH. Anti-VEGF treatment in neovascular age-related macular degeneration: a treat-and-extend protocol over 2 years. Retina. Aug 2014;34(8):1531-1538. [doi: 10.1097/IAE.0000000000000134] [Medline: 24637667]
4. Kang C, Lo JE, Zhang H, Ng SM, Lin JC, Scott IU, et al. Artificial intelligence for diagnosing exudative age-related macular degeneration. Cochrane Database Syst Rev. Oct 17, 2024;10(10):CD015522. [doi: 10.1002/14651858.CD015522.pub2] [Medline: 39417312]
5. Garweg JG, Zirpel JJ, Gerhardt C, Pfister IB. The fate of eyes with wet AMD beyond four years of anti-VEGF therapy. Graefes Arch Clin Exp Ophthalmol. Apr 2018;256(4):823-831. [FREE Full text] [doi: 10.1007/s00417-018-3907-y] [Medline: 29397436]
6. Ho AC, Kleinman DM, Lum FC, Heier JS, Lindstrom RL, Orr SC, et al. Baseline visual acuity at wet AMD diagnosis predicts long-term vision outcomes: an analysis of the IRIS Registry. Ophthalmic Surg Lasers Imaging Retina. Nov 01, 2020;51(11):633-639. [FREE Full text] [doi: 10.3928/23258160-20201104-05] [Medline: 33231696]
7. Lu Y, Huang W, Zhang Y, Huang X, Zhang X, Ma H, et al. Factors for visual acuity improvement after anti-VEGF treatment of wet age-related macular degeneration in China: 12 months follow up. Front Med (Lausanne). Nov 11, 2021;8:735318. [FREE Full text] [doi: 10.3389/fmed.2021.735318] [Medline: 34859005]
8. Skelly A, Bezlyak V, Liew G, Kap E, Sagkriotis A. Treat and extend treatment interval patterns with anti-VEGF therapy in nAMD patients. Vision (Basel). Aug 26, 2019;3(3):41. [FREE Full text] [doi: 10.3390/vision3030041] [Medline: 31735842]
9. Gemenetzi M, Patel PJ. A systematic review of the treat and extend treatment regimen with anti-VEGF agents for neovascular age-related macular degeneration. Ophthalmol Ther. Jun 2017;6(1):79-92. [FREE Full text] [doi: 10.1007/s40123-017-0087-5] [Medline: 28451952]
10. Hassan N. OCT biomarkers in wet type age related macular degeneration. Acta Ophthalmol. Jan 18, 2024;102(S279). [doi: 10.1111/aos.16339]
11. Hanson RL, Airody A, Sivaprasad S, Gale RP. Optical coherence tomography imaging biomarkers associated with neovascular age-related macular degeneration: a systematic review. Eye (Lond). Aug 2023;37(12):2438-2453. [FREE Full text] [doi: 10.1038/s41433-022-02360-4] [Medline: 36526863]
12. Szeto SK, Hui VW, Siu V, Mohamed S, Chan CK, Cheung CY, et al. Recent advances in clinical applications of imaging in retinal diseases. Asia Pac J Ophthalmol (Phila). 2023;12(2):252-263. [FREE Full text] [doi: 10.1097/APO.0000000000000584] [Medline: 36650100]
13. Heinke A, Zhang H, Broniarek K, Michalska-Małecka K, Elsner W, Galang CM, et al. Cross-instrument optical coherence tomography-angiography (OCTA)-based prediction of age-related macular degeneration (AMD) disease activity using artificial intelligence. Sci Rep. Nov 07, 2024;14(1):27085. [FREE Full text] [doi: 10.1038/s41598-024-78327-0] [Medline: 39511248]
14. Schranz M, Bogunovic H, Deak G, Sadeghipour A, Reiter GS, Schmidt-Erfurth U. Linking disease activity with optical coherence tomography angiography in neovascular age related macular degeneration using artificial intelligence. Sci Rep. Aug 20, 2024;14(1):19278. [FREE Full text] [doi: 10.1038/s41598-024-70234-8] [Medline: 39164449]
15. Basilious A, Yuan AT, Sheidow TG. Outcomes of 16-week extension of anti-VEGF therapy in neovascular age-related macular degeneration. Can J Ophthalmol. Apr 2026;61(2):437-442. [FREE Full text] [doi: 10.1016/j.jcjo.2025.10.002] [Medline: 41183785]
16. Most JA, Folk GA, Walker EH, Nagel ID, Mehta NN, Flester E, et al. Evaluating the clinical utility of multimodal large language models for detecting age-related macular degeneration from retinal imaging. Sci Rep. Sep 26, 2025;15(1):33214. [FREE Full text] [doi: 10.1038/s41598-025-18306-1] [Medline: 41006661]
17. Qin Z, Liu Y, Yin Y, Ding J, Zhang H, Li A, et al. LMOD+: a comprehensive multimodal dataset and benchmark for developing and evaluating multimodal large language models in ophthalmology. ACM Trans Comput Healthc. Jun 22, 2026;7(3):1-38. [FREE Full text] [doi: 10.1145/3801746]
18. Shen Y, Chen Q, He X, Agrawal R, Grzybowski A, Jin K, et al. Automated report generation in ophthalmology: integrating artificial intelligence, multimodal imaging, and clinical data. Ophthalmol Ther. Feb 19, 2026;15(3):1001-1020. [doi: 10.1007/s40123-026-01316-1] [Medline: 41714517]

XSL•FO
RenderX

19. Zhao Z, Zhang W, Chen X, Song F, Gunasegaram J, Huang W, et al. Slit lamp report generation and question answering: development and validation of a multimodal transformer model with large language model integration. J Med Internet Res. Dec 30, 2024;26:e54047. [FREE Full text] [doi: 10.2196/54047] [Medline: 39753218]
20. Agbareia R, Omar M, Zloto O, Glicksberg BS, Nadkarni GN, Klang E. Multimodal LLMs for retinal disease diagnosis via OCT: few-shot versus single-shot learning. Ther Adv Ophthalmol. May 20, 2025;17:25158414251340569. [FREE Full text] [doi: 10.1177/25158414251340569] [Medline: 40400723]
21. Zhang Z, Zhang H, Pan Z, Bi Z, Wan Y, Song X, et al. Evaluating large language models in ophthalmology: systematic review. J Med Internet Res. Oct 27, 2025;27:e76947. [FREE Full text] [doi: 10.2196/76947] [Medline: 41144954]
22. Zedadra A, Salah-Salah MY, Zedadra O, Guerrieri A. Multi-modal AI for multi-label retinal disease prediction using OCT and fundus images: a hybrid approach. Sensors (Basel). Jul 19, 2025;25(14):4492. [FREE Full text] [doi: 10.3390/s25144492] [Medline: 40732620]
23. Liu PR, Bansal S, Dinh J, Pawar A, Satishkumar R, Desai S, et al. MedChat: a multi-agent framework for multimodal diagnosis with large language models. In: Proceedings of the 2025 IEEE 8th International Conference on Multimedia Information Processing and Retrieval (MIPR 2025). New York, NY. IEEE; 2025:456-462.
24. Wang W. LLM-based multi-agent system for neuro-ophthalmic diagnosis and personalized treatment planning. Front Neurosci. Oct 06, 2025;19:1688509. [FREE Full text] [doi: 10.3389/fnins.2025.1688509] [Medline: 41122382]
25. Karunanayake N. Next-generation agentic AI for transforming healthcare. Inform Health. Sep 2025;2(2):73-83. [doi: 10.1016/j.infoh.2025.03.001]
26. OpenAI. GPT-4 technical report. arXiv. Preprint posted online on March 15, 2023. [FREE Full text] [doi: 10.48550/arXiv.2303.08774]
27. Wu Q, Bansal G, Zhang J, Wu Y, Li B, Zhu E, et al. AutoGen: enabling next-gen LLM applications via multi-agent conversation. In: Proceedings of the First Conference on Language Modeling (COLM 2024). 2024. Presented at: COLM 2024; Oct 7-9, 2024; Philadelphia, PA. URL: https://openreview.net/forum?id=BAakY1hNKS
28. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems 35. New Orleans, LA. Neural Information Processing Systems Foundation; 2022:24824-24837.
29. Liu J, Wang Y, Du J, Zhou JT, Liu Z. MedCoT: medical chain of thought via hierarchical expert. In: Al-Onaizan Y, Bansal M, Chen YN, editors. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA. Association for Computational Linguistics; 2024:17371-17389.
30. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. Mar 1977;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]
31. Joshi A, Kale S, Chandel S, Pal DK. Likert scale: explored and explained. Curr J Appl Sci Technol. Jan 10, 2015;7(4):396-403. [doi: 10.9734/bjast/2015/14975]

Abbreviations

AMD: age-related macular degeneration
CNV: choroidal neovascularization
LLM: large language model
OCT: optical coherence tomography
OCTA: optical coherence tomography angiography
PED: pigment epithelial detachment
SRF: subretinal fluid
VEGF: vascular endothelial growth factor

Edited by J Sarvestan; submitted 10.Jul.2026; peer-reviewed by X Zhan; comments to author 14.Sep.2026; revised version received 18.Sep.2026; accepted 02.Nov.2026; published 23.Nov.2026

Please cite as:
Jiang X, Tao J, Yu C, Wang L, Fang X, Xue W, Qin Q
A Multiagent Large Language Model Framework for Longitudinal Multimodal Tracking of Wet Age-Related Macular Degeneration: Retrospective Feasibility Study
JMIR Form Res 2026;10:e106685
URL: https://formative.jmir.org/2026/1/e106685
doi: 10.2196/106685
PMID:

XSL•FO
RenderX

©Xuan Jiang, Jinhua Tao, Chunxia Yu, Leilei Wang, Xiaoling Fang, Wenwen Xue, Qin Qin. Originally published in JMIR Formative Research (https://formative.jmir.org), 23.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Formative Research, is properly cited. The complete bibliographic information, a link to the original publication on https://formative.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX