

An agentic multidisciplinary framework for diabetic retinopathy treatment planning

Tingting Shen

Southern University of Science and Technology

Zhiyuan Cai

Southern University of Science and Technology

Pujin Cheng

Southern University of Science and Technology

Li Lin

Southern University of Science and Technology

Meng Tian

Beijing Tongren Hospital

Honghui Xia

Second Affiliated Hospital of Guangzhou Medical University

Zhihao Jiang

Zhongshan Ophthalmic Center, Sun Yat-sen University

Ting Zhang

Zhongshan Ophthalmic Center, Sun Yat-sen University

Yantao Wei

Zhongshan Ophthalmic Center, Sun Yat-sen University

Zhaoyang Wang

Beijing Tongren Hospital

Jin Yuan

Beijing Tongren Hospital

Wenjun Xu

Peng Cheng Laboratory

Xiaoying Tang

tangxy@sustech.edu.cn

Southern University of Science and Technology

Article

Keywords:

Posted Date: August 5th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10335654/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

An agentic multidisciplinary framework for diabetic retinopathy treatment planning

Tingting Shen^{1,2}, Zhiyuan Cai¹, Pujin Cheng¹, Li Lin¹, Meng Tian³,
Honghui Xia⁴, Zhihao Jiang⁵, Ting Zhang⁵, Yantao Wei⁵, Zhaoyang Wang³,
Jin Yuan³, Wenjun Xu^{2*}, Xiaoying Tang^{1*}

¹Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen, China.

²Peng Cheng Laboratory, Shenzhen, China.

³Beijing Tongren Eye Center, Beijing Tongren Hospital, Capital Medical University, Beijing, China.

⁴Department of Ophthalmology, Gaoyao Hospital, Second Affiliated Hospital of Guangzhou Medical University, Zhaoqing, China.

⁵State Key Laboratory of Ophthalmology, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, 510060, China.

*Corresponding author(s). E-mail(s): xuwj@pcl.ac.cn; tangxy@sustech.edu.cn;

Abstract

Diabetic retinopathy (DR) is a common cause of vision loss among working-age adults, and DR treatment planning requires multidisciplinary collaboration across ophthalmology and endocrinology. In routine practice, limited specialist resources, documentation burden, and the difficulty of implementing multidisciplinary team consultation prevent many patients from receiving individualized treatment plans. We present three linked components: a large-scale, publicly available DR treatment-planning dataset, linking fundus images, clinical records, and treatment plans; DRAgent, a role-based multi-agent system in which clinical agents collaborate to generate individualized plans; and an evaluation framework combining guideline compliance, objective metrics, and human evaluation. DRAgent outperforms the corresponding standalone LLMs on nearly all metrics, raising human evaluation scores from 4.02 to 4.80/5. Ablation studies show that removing any specialist agent degrades performance in its clinical domain, confirming complementary roles. By turning multidisciplinary decision-making into an evaluable system, DRAgent may provide a useful starting point for developing similar systems in other diseases requiring cross-specialty care.

1 Introduction

Diabetic retinopathy (DR) is a major cause of vision loss among working-age adults, yet much of this loss can be prevented through timely treatment. However, DR treatment planning requires multidisciplinary collaboration between ophthalmology and endocrinology, and complex cases often require input from senior specialists. This requirement is particularly difficult to meet in primary and

resource-limited settings, where many patients cannot receive standardized treatment in a timely manner. Even in higher-tier hospitals where specialist resources are concentrated, clinicians face a substantial documentation burden. Previous research shows that, in an academic ophthalmology practice, documentation accounts for approximately 38% of the total time spent per patient encounter [1]. Together, these constraints limit the efficiency and accessibility of DR care, leaving some patients to miss the optimal treatment window.

Large language models (LLMs) offer a new route to closing this gap [2, 3]. Med-PaLM, developed by Google Health, was the first AI system to reach a passing score on USMLE-style questions [3, 4]; leading models now score 84–90% on such examinations, approaching or exceeding the average physician [5], and exceed 90% diagnostic accuracy on common clinical scenarios [6]. However, this competence does not carry over to practice. Across 39 benchmarks, performance falls from 84-90% on knowledge-based tests to 45-69% on practice-based tasks [5], and in an evaluation of 21 frontier LLMs, differential diagnosis failed in over 80% of cases even though final diagnoses were largely correct [7]. Treatment planning sits further downstream than diagnosis and has been less studied [8, 9].

This limitation stems primarily from three key challenges. **1) Data scarcity:** publicly available medical datasets predominantly focus on diagnostic labels, while high-quality, structured treatment-plan data remain limited, restricting the development and validation of treatment-planning models. **2) Decision complexity:** treatment-plan generation requires the integration of patient-specific characteristics, guideline recommendations, and clinical feasibility, making it substantially more complex than conventional diagnostic tasks [8]. **3) Evaluation difficulty:** treatment plans are inherently open-ended and not fully standardized, and no two clinicians write them alike. Without a systematic evaluation framework, the clinical quality of AI-generated plans cannot be reliably assessed. Therefore, establishing appropriate datasets, methodological frameworks, and evaluation strategies to support intelligent treatment planning remains a major challenge in medical AI.

Multi-agent systems (MAS) provide a promising pathway for addressing these challenges. Unlike a standalone LLM, which generates text in response to a prompt, an agent is defined by an assigned role, access to knowledge beyond its parameters, the capacity to plan and act within that role, and the ability to collaborate with other agents toward a shared goal [10, 11]. Such role-based collaboration among specialized agents approximates the way multidisciplinary clinical decisions are made in practice. Frameworks such as MedAgents [12], MDAgents [8], and conversational multi-agent consultation systems [13] demonstrate that simulating cross-specialty expert collaboration can improve diagnostic accuracy in complex cases. Agentic systems such as DeepRare further show that LLM-powered agents can integrate heterogeneous clinical inputs, specialized tools, and traceable medical evidence to support complex diagnostic workflows [14]. Recent work evaluates multi-agent LLMs using real multidisciplinary team meeting data, showing that role-specialized, conversation-based agents can better align with expert consensus in complex cancer treatment decisions [15]. Meanwhile, Agent Hospital [16] and AI Hospital [17] further explore the self-evolving capabilities of agents in simulated healthcare environments. However, existing medical MAS studies remain primarily focused on diagnostic tasks and do not yet sufficiently address treatment planning. The few treatment-related studies available often simplify the output, for example by using probability fusion, predefined severity categories, or single surrogate indicators, rather than generating complete individualized treatment plans [8, 9]. Meanwhile, existing medical AI evaluation frameworks are largely built around question-answering correctness or diagnostic accuracy. Mainstream benchmarks such as MedQA [18], MedMCQA [19], and PubMedQA [20] are centered on knowledge-based question answering; even recent treatment-oriented evaluation efforts often reduce assessment to correctness judgments [21], making it difficult to comprehensively evaluate the clinical rationality, safety, and individual suitability of treatment plans. In ophthalmology specifically, recent advances in multimodal ophthalmic artificial intelligence improve image understanding, disease grading, and medical language interaction [22]. Representative models such as EyeCLIP [23], EyecareGPT [24], and Ophtha-LLaMA2 [25] advance ophthalmic image understanding and medical language interaction. Nevertheless, multi-agent treatment-planning systems that cover the full treatment workflow, integrate multidisciplinary collaboration, and remain grounded in clinical guidelines are still at an exploratory stage [26].

This study presents an agentic multidisciplinary framework for DR treatment planning, designed to alleviate specialist resource shortages, documentation burden, and the limited scalability of multidisciplinary collaboration. The framework comprises three linked components. First, we construct a large-scale, publicly available DR treatment-planning dataset, linking fundus images, structured clinical records, and standardized treatment plans to provide a data foundation for treatment-planning research. Second, we build **DRAgent**, a clinically role-based multi-agent system in which an ophthalmologist, a laser surgeon, an endocrinologist, and a pharmacist collaborate to generate individualized, guideline-compliant treatment plans. Third, we establish a multidimensional evaluation framework integrating guideline compliance, objective metrics, and human evaluation. Together, these components enable the systematic generation and evaluation of DR treatment plans.

2 Results

This section reports the DR treatment-planning dataset and the experimental results. The experiments consist of two parts: comparative experiments, which compare the performance of different backbone large language models when generating treatment plans independently versus when embedded within the DRAgent; and ablation experiments, which evaluate the contribution of each specialist agent in DRAgent to the quality of the final treatment plan. We select several representative models as baselines, including closed-source general-purpose models GPT-5 and Gemini 2.5 Flash, open-source general-purpose models Qwen3.5-flash and DeepSeek-V3.1, and the medical-domain-specific model HuatuoGPT-3. Detailed descriptions of these models are provided in Supplementary Note 1.

2.1 DR treatment-planning dataset

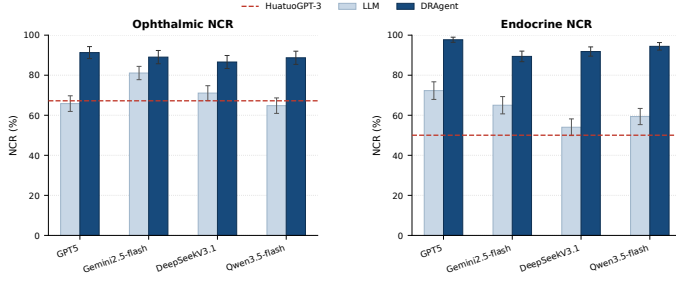
A contribution of this study is the construction of a large-scale DR treatment-planning dataset. This dataset integrates fundus images, structured clinical records, and standardized treatment plans, covering four disease stages from mild non-proliferative diabetic retinopathy (NPDR) to proliferative diabetic retinopathy (PDR) and comprising approximately 10,000 cases in total. It consists of three components. First, 100 multi-center real-world clinical cases collected from Zhongshan Ophthalmic Center (ZOC) of Sun Yat-sen University, a leading ophthalmic specialty center in China, and Gaoyao District People’s Hospital in Zhaoqing. Second, 1,040 hybrid real-synthetic cases, in which clinical records and treatment plans are generated from real fundus images and subsequently reviewed and revised by associate-senior or higher-level ophthalmologists from Beijing Tongren Hospital until they meet clinical standards. Third, 8,908 large-scale automatically generated cases produced by DRAgent. The real-world clinical subset and the hybrid real-synthetic subset serve as the primary validation datasets in this study. Through an “image-to-report-to-plan” construction pathway, the dataset provides case-level clinical context for treatment-plan generation and supports the evaluation of treatment-planning systems. The large-scale generated subset is mainly intended for future model training, fine-tuning, and related research.

2.2 Comparative Experiments

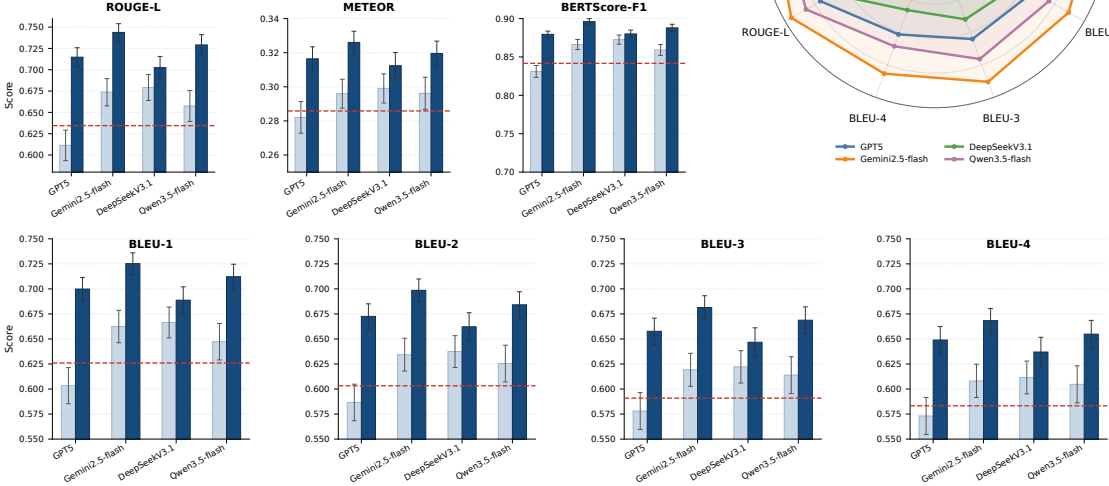
2.2.1 Guideline Compliance

Clinical guidelines constitute the standardized compliance reference for DR treatment. In this study, we adopt the Normalized Compliance Rate (NCR) to evaluate the consistency between generated treatment plans and guideline recommendations, thereby quantifying their clinical rationality and standardization. NCR is computed separately for the ophthalmic treatment and endocrine management components of each plan. Experiments are conducted on 588 hybrid real-synthetic cases sampled from the 1,040-case hybrid subset and 100 real clinical cases, with results reported separately for the two datasets. The former consists of physician-revised reference cases, providing a relatively large sample size and a more balanced distribution across disease stages; the latter is derived from authentic multi-center clinical cases and is primarily used to assess the model’s generalizability and robustness in real-world clinical settings.

A. Guideline compliance assessment



B. Objective similarity assessment



C. DRAgent backbone profile

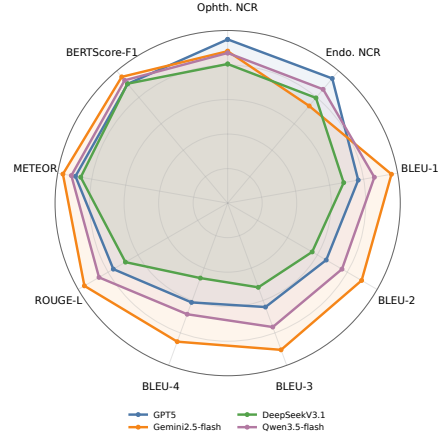


Fig. 1 Comparative performance of DRAgent against standalone LLMs and HuatuoGPT-3 on the hybrid real-synthetic set. (A) Guideline-compliance NCR for the ophthalmic and endocrine components, for HuatuoGPT-3, each backbone LLM and its DRAgent counterpart, with 95% CI error bars. (B) Reference-based objective similarity for HuatuoGPT-3, each backbone LLM and its DRAgent counterpart, with 95% CI error bars. (C) Radar comparison of the DRAgent objective profile across the four backbones.

Hybrid Real-Synthetic Data

On the hybrid real-synthetic evaluation dataset, DRAgent achieves higher NCR values than the standalone LLM baselines in both the ophthalmic treatment and endocrine management domains across all four backbone model architectures (Fig. 1A; full numerical values in Supplementary Table 4). For the ophthalmic component, standalone LLMs are generally able to generate broadly reasonable ophthalmic recommendations; however, in some cases, they still exhibit deviations in key treatment decisions, which may pose potential risks in clinical application. In addition, single LLMs tend to omit or incompletely specify procedural details. By contrast, DRAgent systematically addresses these limitations through its multi-expert consultation mechanism: the laser surgeon discusses unreasonable or insufficient aspects of the treatment plan with the ophthalmologist, thereby further refining ophthalmic decision-making. For the endocrine management component, the improvement achieved by DRAgent is more pronounced, which is consistent with the cross-disciplinary nature of this task. Because the input information is primarily centered on ophthalmic findings, standalone LLMs often fail to proactively generate comprehensive metabolic management plans. In contrast, DRAgent explicitly incorporates an endocrinologist and a pharmacist, thereby routing each case through systematic metabolic management and medication-safety assessment, which substantially improves guideline adherence in the endocrine component. The consistency of the results across all four backbone models indicates that the performance gains are primarily attributable to the multi-agent collaboration itself, rather than to the capability differences of any specific backbone model.

Table 1 Center-stratified guideline compliance NCR on real clinical data.

Center	Model	Ophth. LLM	Ophth. DRAgent	Endo. LLM	Endo. DRAgent
ZOC	GPT-5	76.16	92.36	35.49	63.63
	Gemini 2.5 Flash	80.60	86.27	36.03	53.16
	DeepSeek-V3.1	63.00	79.51	15.26	35.35
	Qwen3.5-flash	67.96	74.70	30.69	35.20
Gaoyao	GPT-5	72.95	86.05	36.06	60.74
	Gemini 2.5 Flash	76.13	86.94	28.13	58.59
	DeepSeek-V3.1	56.63	83.50	7.07	47.72
	Qwen3.5-flash	64.55	86.70	27.80	43.53

Real Clinical Data

Real clinical cases are used as an independent validation source to further assess model performance. As shown in Table 1, DRAgent achieves consistent performance improvements over the standalone LLM baselines at both ZOC and Gaoyao, indicating that the guideline-adherence advantage observed in the controlled hybrid benchmark can be transferred to authentic clinical environments. In addition, compared with LLMs, treatment plans generated by DRAgent exhibit more stable guideline compliance across different medical centers and DR severity grades. Although ZOC and Gaoyao differ in patient severity distribution and local documentation practices, both centers show the same direction of improvement, further supporting the generalizability and robustness of DRAgent across diverse clinical settings.

2.2.2 Objective Evaluation

We further evaluate the objective similarity between the generated treatment plans and the reference plans. The evaluation datasets are consistent with those used for the guideline-adherence analysis, comprising 100 real clinical cases and 588 hybrid real-synthetic cases. In the hybrid real-synthetic evaluation set, physician-revised treatment plans are used as the reference standard, whereas in the real clinical case set, authentic physician-authored treatment plans serve as the reference standard. Because multiple large language models are incorporated during the generation stage of the hybrid real-synthetic dataset, this design effectively reduces the potential bias introduced by the generation style of any single model. Specifically, we compare the lexical overlap and semantic similarity between model-generated treatment plans and reference plans. The evaluation metrics include BLEU-1 to BLEU-4, ROUGE-L, METEOR, and BERTScore-F1. To ensure that the assessment focuses on core therapeutic decisions, all treatment-plan texts are first processed through structured summarization, and the final metrics are aggregated using weighted averaging according to the clinical importance of different treatment-plan components. The detailed evaluation procedure is described in Section 4.3.2.

Validation on Hybrid Real-Synthetic Data

The results are shown in Fig. 1B, with the full numerical values in Supplementary Table 5. Across all four backbone architectures, DRAgent outperforms the corresponding standalone LLMs on all objective similarity metrics. Improvements in lexical-overlap metrics, including BLEU and ROUGE-L, indicate that the treatment plans generated by DRAgent achieve greater textual concordance with expert-revised reference plans, suggesting that the specialist consultation process guides the model toward a more clinically appropriate terminology system and plan structure. The concurrent improvement in BERTScore-F1 further indicates that DRAgent enhances not only surface-level textual matching but also the alignment with the clinical semantics and therapeutic strategies embedded in the reference plans. Meanwhile, HuatuoGPT-3 does not consistently outperform the general-purpose LLMs across the evaluated metrics. This suggests that, under the current evaluation setting, medical-domain specialization alone does not translate into a clear advantage for DR treatment-plan generation.

Furthermore, Fig. 1C provides an integrated overview of DRAgent performance across guideline adherence and objective similarity metrics using different backbone models. Overall, DRAgent based

Table 2 Center-stratified objective similarity on real clinical data.

Center	Model	BLEU-1		ROUGE-L		METEOR		BERTScore-F1	
		LLM	DRAgent	LLM	DRAgent	LLM	DRAgent	LLM	DRAgent
ZOC	GPT-5	0.550	0.659	0.569	0.678	0.238	0.276	0.823	0.874
	Gemini 2.5 Flash	0.592	0.646	0.611	0.672	0.238	0.264	0.834	0.871
	DeepSeek-V3.1	0.600	0.622	0.617	0.642	0.265	0.267	0.844	0.853
	Qwen3.5-flash	0.628	0.623	0.650	0.648	0.288	0.297	0.854	0.856
Gaoyao	GPT-5	0.551	0.637	0.568	0.649	0.227	0.254	0.808	0.833
	Gemini 2.5 Flash	0.572	0.624	0.590	0.643	0.269	0.252	0.820	0.840
	DeepSeek-V3.1	0.543	0.650	0.556	0.673	0.267	0.270	0.808	0.850
	Qwen3.5-flash	0.576	0.582	0.592	0.607	0.310	0.276	0.819	0.819

on Gemini 2.5 Flash achieves the best overall performance, followed by Qwen3.5-flash and GPT-5, whereas GPT-5 achieves the highest guideline compliance; DeepSeek-V3.1 shows relatively lower performance. No single backbone leads on all axes; the differences among backbone models may be associated with variations in their intrinsic reasoning capabilities and generation preferences.

Validation on Real Clinical Data

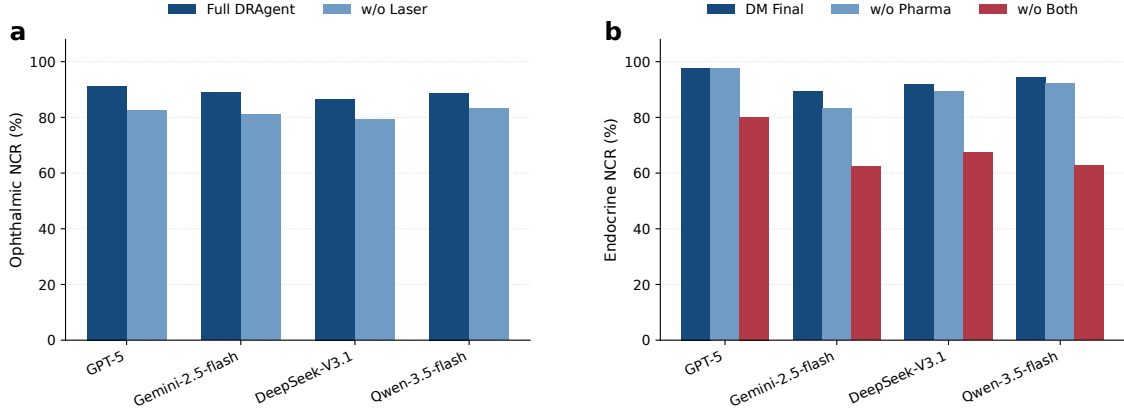
To determine whether the improvements can be transferred to real clinical environments, we further perform independent validation on real clinical datasets. Results from both medical centers consistently demonstrate that DRAgent outperforms the corresponding standalone LLMs. Reversed trends are observed only in a small number of metrics, mainly in the results for Qwen3.5-flash and Gemini 2.5 Flash, which may be attributed to differences in the intrinsic generation capability of the backbone model and the distribution characteristics of real clinical narratives. Further analysis reveals that the magnitude of improvement achieved by DRAgent is associated with the baseline capability of the backbone model and case characteristics. Standalone LLMs with lower initial similarity benefit more substantially from DRAgent, whereas models that already exhibit higher textual consistency have relatively limited room for further improvement because their generated outputs are already closer to the reference plans. These lexical and semantic improvements are consistently maintained across authentic clinical cases from two independent medical centers, further demonstrating that the multi-agent collaboration can effectively enhance the quality of treatment-plan generation (Table 2).

2.2.3 Human Evaluation

Human evaluation follows a dual blinded review and expert adjudication protocol. The evaluation set consists of 187 hybrid real-synthetic cases selected from the previous evaluation dataset. For each case, the standalone LLM and DRAgent independently generate treatment plans based on the same patient information, which are randomly assigned as Plan A and Plan B. Two ophthalmologists then independently review each paired plan under blinded conditions and score them according to a five-point clinical rationality rating scale. The initial dual-review assessment demonstrates that, across all four backbone model architectures, treatment plans generated by DRAgent achieve higher clinical rationality scores than the corresponding LLM baselines (Table 3). Among 748 paired comparisons (4 models $\times$ 187 cases), the inter-reviewer agreement rate ranges from 76.6% to 86.7%, with Cohen’s kappa values ranging from 0.644 to 0.790, indicating good consistency between reviewers in the initial clinical assessments. For 140 paired comparisons (18.7%) with disagreement in relative preference, a third ophthalmologist independently adjudicates the cases, and the final scores are determined based on the adjudicated consensus. The final adjudicated results further validate the initial findings: across all four backbone architectures, DRAgent consistently achieves higher clinical rationality scores than the corresponding LLM baselines (Table 3). All paired comparisons reach statistical significance, demonstrating a stable advantage of DRAgent-generated treatment plans in terms of clinical rationality. The change in the score difference between DRAgent and LLM before and after adjudication is not statistically significant ($p = 0.396$), indicating that third-expert adjudication resolves individual reviewer disagreements without altering the overall conclusion that DRAgent outperforms standalone LLMs. Furthermore, we perform two-way analysis of variance (ANOVA), with backbone

Table 3 Human evaluation. Significance: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Model	Reviewer 1		Reviewer 2		Final		Cohen’s κ	Sig.
	LLM	DRAgent	LLM	DRAgent	LLM	DRAgent		
GPT-5	4.335	4.883	4.043	4.862	4.245	4.904	0.741	***
Gemini 2.5 Flash	4.415	4.957	4.351	4.915	4.393	4.933	0.790	***
DeepSeek-V3.1	3.266	4.617	3.282	4.564	3.251	4.586	0.679	***
Qwen3.5-flash	4.410	4.793	3.798	4.676	4.181	4.780	0.644	***
Mean	4.11	4.81	3.87	4.75	4.02	4.80	—	—

**Fig. 2** Agent ablation across four backbone-specific DRAgent systems. **a**, Ophthalmic NCR with and without the laser surgeon. **b**, Endocrine NCR with and without the pharmacist, and without both the pharmacist and the endocrinologist.

model type and DR severity grade as influencing factors, to evaluate their effects on human evaluation scores. The results show that both factors significantly affect clinical ratings; however, their effects are substantially stronger for the standalone LLMs (backbone model, $F = 96.50$; DR grade, $F = 108.31$) than for DRAgent (backbone model, $F = 18.96$; DR grade, $F = 3.41$), indicating that standalone LLM performance is more sensitive to model selection and disease severity. In contrast, DRAgent maintains consistently higher and more stable clinical rationality scores across different backbone architectures and DR severity grades, further supporting the robustness of the multi-agent collaborative workflow in complex treatment-plan generation tasks.

Clinically, DRAgent plans receive final ratings between 4 and 5 (mean 4.80), placing them between “largely reasonable, requiring only minor revision” and “highly reasonable, directly usable for clinical reference”.

2.3 Ablation Studies

To evaluate the contribution of individual agents to the quality of the final treatment plans, we perform ablation analyses on the 588-case hybrid real-synthetic evaluation set, using guideline adherence as the primary evaluation metric. Since different specialist roles contribute to distinct clinical components, the ophthalmologist and laser surgeon are evaluated based on ophthalmic treatment adherence, whereas the endocrinologist and pharmacist are evaluated based on endocrine management adherence. Three ablation configurations are designed to investigate the impact of individual expert roles on treatment-plan quality; the full numerical values are reported in Supplementary Table 6.

Laser surgeon

We evaluate the contribution of the laser surgeon using the ophthalmic guideline adherence assessment framework (Fig. 2a). Across all four backbone architectures, removing the laser surgeon results in decreased ophthalmic guideline adherence, indicating that this specialist role plays a critical role

in improving the completeness and standardization of ophthalmic treatment plans. Without this specialized input, generated plans can generally capture the overall diagnostic and therapeutic direction but show clear deficiencies in surgical planning details.

Pharmacist and Endocrinologist

We evaluate the contributions of the endocrinologist and pharmacist using the endocrine guideline adherence framework (Fig. 2b). Removing the pharmacist results in a modest decrease in adherence scores, suggesting that the endocrinologist is able to address most systemic metabolic management requirements, whereas the pharmacist primarily provides complementary optimization in medication selection, dosage adjustment, and medication-related considerations. In contrast, simultaneous removal of both the endocrinologist and pharmacist leads to a substantial decline in adherence, demonstrating that an ophthalmology-centered system alone cannot adequately address the systemic management needs of patients with DR. The endocrinologist primarily drives metabolic management, while the pharmacist further enhances plan completeness through medication optimization.

Collectively, the ablation results demonstrate that different specialist agents contribute distinct and complementary functions during treatment-plan generation. The laser surgeon primarily improves ophthalmic procedural planning, whereas the endocrinologist and pharmacist jointly ensure comprehensive systemic metabolic management. Removal of any key specialist role results in reduced guideline adherence, further supporting the importance of the multi-expert collaboration mechanism for complex DR treatment planning.

3 Discussion

In this study, we present an agentic multidisciplinary framework for DR treatment planning, comprising a large-scale treatment-planning dataset, DRAgent, and a multidimensional evaluation framework. DRAgent simulates real-world multidisciplinary decision-making, in which clinical agents holding distinct roles reason within their specialty and negotiate a shared plan. Systematic validation demonstrates that, compared with standalone LLMs, DRAgent achieves consistent improvements in guideline compliance, objective metrics, and human evaluation. These findings suggest that structured multi-expert collaboration can effectively enhance the clinical applicability of LLMs in complex treatment-planning tasks.

Unlike conventional text generation tasks, treatment planning requires the integration of multidimensional patient information and the coordination of decisions across different clinical specialties. Although LLMs show substantial potential in the medical domain, they continue to face challenges in complex clinical decision-making. DRAgent addresses this limitation by assigning clinical roles to separate agents, allowing multidisciplinary perspectives to jointly contribute to the final treatment plan. The consistent improvements observed across different backbone architectures indicate that the advantage of DRAgent is not solely attributable to the capabilities of the underlying language models, but more importantly to the discussion-based collaboration among clinical agents.

This study also highlights the importance of establishing evaluation strategies for AI-generated treatment plans. Unlike conventional medical AI tasks with relatively well-defined target outputs, treatment plans admit many equally valid formulations, and text similarity alone cannot capture their clinical quality. We combine guideline adherence, objective metrics, and human evaluation, which together assess clinical standardization, agreement with expert references, and practical usability. The backbone comparison underscores why multiple evaluation axes are necessary. No backbone led on all of them: the model whose plans aligned most closely with the references was not the one that best satisfied guideline criteria. Because these axes capture different properties of a treatment plan, no single metric is sufficient to judge its quality.

We further observe that some outputs generated in the current task do not fully adhere to the knowledge-base information and instead rely on the model’s own parametric knowledge. This phenomenon may be related to hallucination and knowledge-boundary issues in large language models: when parametric knowledge conflicts with external knowledge sources, the model may defer to its parametric knowledge, leading to accurate retrieval but deviating outputs [27]. Therefore, in specialized clinical scenarios, model selection should consider not only the intrinsic capability of the

backbone model but also its ability to effectively utilize domain-specific knowledge augmentation strategies.

Ablation experiments further clarify the contributions of different agents within DRAgent. Removal of specific specialist roles leads to performance degradation in the corresponding clinical domains, indicating that the agents provide complementary rather than redundant functions. This supports translating real-world multidisciplinary collaboration into an agent-based system, and offers a design reference for AI systems addressing complex medical decisions.

The clinical value of DRAgent extends beyond treatment-plan generation to its potential role in supporting new models of DR care delivery. In primary and resource-limited care, it offers clinicians a guideline-aligned reference plan where specialist input is unavailable. In specialist ophthalmic clinics, it drafts a structured plan that clinicians revise rather than compose from scratch, reducing documentation burden. For complex cases, it brings multiple specialties to bear on a single plan without convening a multidisciplinary team (MDT) meeting.

This study has several limitations. First, the dataset and knowledge base are derived from Chinese clinical guidelines and practice patterns, and their generalizability to other healthcare systems requires further validation. Second, to avoid potential bias from unverified sources, we draw only on formally published guidelines; this also limits the capacity of the knowledge base for dynamic updating. Third, the present dataset relies on fundus images as the sole imaging modality, as no multimodal DR dataset was available at the time of construction; we plan to collect a multimodal dataset in future work to incorporate additional modalities, which would provide richer diagnostic context and support more precise treatment planning. Fourth, human evaluation was conducted by ophthalmologists. Although ophthalmologists assess DR treatment plans in their entirety, their judgment of the endocrine components does not carry the specificity of an endocrinologist review.

Future work will incorporate endocrinologist review and multimodal inputs, pursue multi-center validation, develop more adaptive agent collaboration strategies, and exploit the large-scale treatment-planning dataset for model training and fine-tuning, with the aim of advancing more efficient and clinically applicable DR treatment-planning systems. Extending this framework to other chronic diseases requiring multidisciplinary collaboration is a further important direction.

4 Methods

This section presents the overall study design, including the construction of the DR treatment-planning dataset, the multi-agent treatment-planning system, and the multidimensional evaluation framework.

4.1 DR treatment-planning Dataset

The DR treatment-planning dataset is a publicly available, large-scale dataset for DR treatment planning, covering three modalities: fundus images, clinical records, and corresponding treatment plans (Fig. 3). In total, the dataset comprises approximately 10,000 cases, including 100 multi-center real-world clinical cases, 1,040 hybrid real-synthetic cases, and 8,908 large-scale automatically generated cases. Disease grading follows the international diagnostic standard for DR, categorizing cases into mild NPDR, moderate NPDR, severe NPDR, or PDR.

4.1.1 Real Data

The real-world clinical subset includes 100 multi-center cases from the ZOC of Sun Yat-sen University and Gaoyao District People’s Hospital in Zhaoqing (50 cases each), comprising mild NPDR ($n = 1$), moderate NPDR ($n = 14$), severe NPDR ($n = 20$), and PDR ($n = 65$). Each case contains a clinical record and the corresponding treatment plan; a subset is accompanied by multimodal ophthalmic imaging, including color fundus photography, optical coherence tomography (OCT), and fundus fluorescein angiography (FFA). All data undergo de-identification and standardization before analysis: identifying information such as patient names and admission numbers is removed, and the text is converted into a structured format following a unified clinical-record and treatment-plan template.

This subset is markedly skewed toward advanced disease, reflecting real-world care-seeking patterns: early DR causes few symptoms, so patients seldom present until vision has deteriorated, and the skew is further amplified along the hospital hierarchy. This distribution indicates that many patients are not reached in time—the unmet need this study addresses.

4.1.2 Hybrid Real-Synthetic Dataset

Because real treatment data are scarce, we construct the hybrid real-synthetic dataset with a “real image + generated text + physician revision” strategy. Fundus images with DR grading labels are obtained from the “Good” quality subset of the EyeQ dataset (derived from EyePACS) [28, 29], and an assistive diagnostic report is generated for each image using EyeGraph (manuscript in preparation). The fundus image, DR grade, and diagnostic report are then jointly processed by a fine-tuned ophthalmology LLM to produce an initial clinical record.

To align the linguistic style of the generated records with authentic clinical writing, we construct a DR-specific terminology mapping library. A language model first drafts examination reports for 40 representative fundus images spanning all DR stages; senior ophthalmologists then revise the wording, and the corrected expressions are consolidated into a library mapping model-generated phrasing to standardized clinical terminology and sentence patterns. This library guides terminology and phrasing during subsequent generation, so that the resulting records follow professional clinical conventions.

The generated clinical records are then distributed evenly across four large language models (GPT-5, Gemini 2.5 Flash, DeepSeek-V3.1, and Qwen3.5-flash), each of which drafts treatment plans from a unified prompt. The four models used for generation are the same four later evaluated. This is deliberate: distributing generation across models prevents any single model’s style from dominating the reference plans, which preserves the neutrality of the subsequent comparisons. Senior ophthalmologists (associate chief physician or above) then review and revise every clinical record and treatment plan through a dedicated annotation system, until each meets real-world clinical standards.

Both templates are derived from clinical practice. The clinical note template follows the outpatient record format used at ZOC, comprising demographic information, chief complaint, history of present illness, past medical history, ophthalmic examination, and preliminary diagnosis. Fields are populated according to what the image can support: findings that are directly observable on fundus imaging (e.g., retinal hemorrhage, exudation, microaneurysms) are derived from image analysis, whereas fields that cannot be inferred from the image, such as symptoms and medical history, are completed by the language model from contextual cues and checked for clinical plausibility. The treatment plan template, finalized through physician discussion, comprises disease diagnosis, therapeutic principles, treatment strategy (medication or surgical intervention), follow-up plan, patient education, and precautions.

Finally, the hybrid real-synthetic dataset spans the full spectrum of disease severity, comprising mild NPDR ($n = 338$), moderate NPDR ($n = 396$), severe NPDR ($n = 222$), and PDR ($n = 84$), for a total of 1,040 cases. All records and treatment plans are revised by clinicians to meet clinical standards, and the dataset serves as the primary evaluation benchmark of this study.

4.1.3 Large-Scale Synthetic Dataset

The architecture of the large-scale dataset maintains consistency with the hybrid real-synthetic dataset. The fundus images are drawn from publicly available datasets, including OIA-DDR [30], MMRDR [31], IDRiD [32], and APTOS 2019 [33], each of which provides fundus images together with DR severity grades on a 0 to 4 scale. We first exclude non-DR (grade 0) cases. For datasets containing more than 1,000 cases, we further apply QuickQual [34] for image-quality screening, retaining the images with the highest predicted probability of “Good or Usable” quality, thereby safeguarding the overall quality of the data. After this curation, the retained images comprise mild NPDR ($n = 2,406$), moderate NPDR ($n = 5,203$), severe NPDR ($n = 771$), and PDR ($n = 528$), for a total of 8,908 cases. The generation process for clinical records follows the protocol established in the hybrid phase, while treatment plans are automatically generated using DRAgent, whose reliability

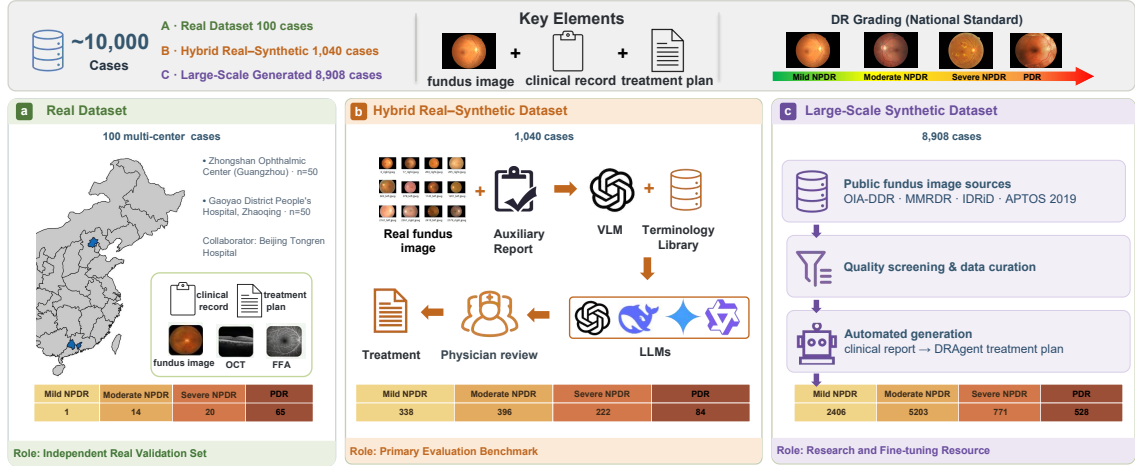


Fig. 3 Overview of the DR treatment-planning dataset, including real clinical data, hybrid real-synthetic data, and large-scale generated data.

is established in the Results section. By integrating public fundus image repositories with an automated generation pipeline, this study constructs a large-scale treatment-plan dataset that provides a data foundation for subsequent research.

In summary, the integrated pipeline yields a publicly available dataset that jointly provides fundus images, clinical records, and treatment plans across the full spectrum of DR severity: real clinical cases for independent validation, hybrid real-synthetic cases as the primary evaluation benchmark, and a large-scale automatically generated cohort for future model training and fine-tuning. By filling the gap in ophthalmic treatment-planning datasets, it supports downstream tasks such as treatment-plan generation and the benchmarking of clinical decision-support systems.

4.2 Multi-Agent Treatment Planning System

We develop a multi-agent treatment-planning system on the DIFY platform. As shown in Fig. 4, the system simulates the clinical treatment-planning process for DR patients through a multi-agent collaborative mechanism.

4.2.1 Clinical Role Simulation

To reproduce the structure of MDT management, DRAgent incorporates diverse clinical expert roles, employing LLM-based agents to simulate dynamic medical inquiries and collaborative consultation processes. Given the multifaceted nature of DR, we integrate two primary disciplines, ophthalmology and endocrinology, to establish key clinical personas: an outpatient physician, an ophthalmologist, a laser surgeon, an endocrinologist, a pharmacist, and a nurse. Specifically, the ophthalmologist and laser surgeon belong to the ophthalmology department; the endocrinologist and pharmacist represent the endocrinology department; the outpatient physician and nurse handle intake and consolidation.

The division of labor is defined. The outpatient physician collects patient data and produces the clinical record. The ophthalmologist, acting as attending physician, drafts the initial treatment plan and leads the interdisciplinary discussion. The other specialists contribute judgements within their own domain and deliberate with the ophthalmologist. The nurse consolidates the outcome of each discussion into an updated plan, which keeps the structure of the plan stable across successive rounds.

4.2.2 Knowledge-Driven Single-Agent Architecture

The outpatient physician and nurse are assigned GPT-5 with respective prompts, while the other agents are built on LLMs and augmented with specialized knowledge bases and custom prompts.

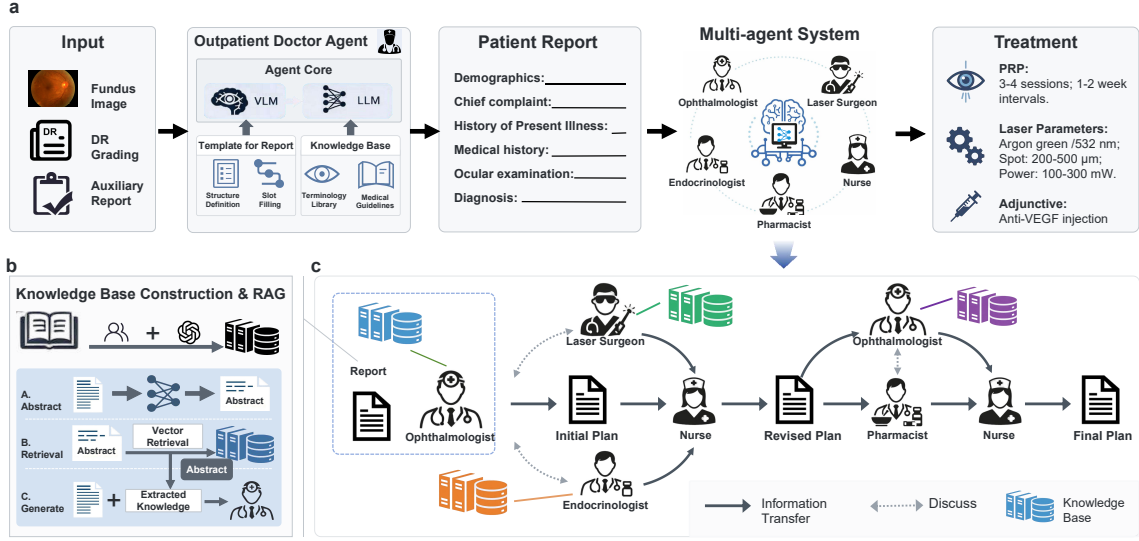


Fig. 4 Overview of the DRAgent multi-agent treatment-planning system. (a) Automated pipeline: from fundus images to clinical records and final treatment plans. (b) Construction of clinical knowledge bases and the summarization-based RAG framework. (c) Multi-agent collaboration system.

All agent prompts are listed in Supplementary Note 2. We evaluate four backbone models—GPT-5 [35], Gemini 2.5 Flash [36], DeepSeek-V3.1 [37] and Qwen3.5-flash [38]—together with the medical-domain model HuatuoGPT-3 [39] as a standalone baseline; Claude Opus 4.8 [40] is used solely as the third-party scorer. The exact model version strings are listed in Supplementary Note 1.

Knowledge Base Construction Retrieval-augmented generation (RAG) is widely used to ground language-model outputs in external knowledge sources and shows potential in clinical medicine [41]. DRAgent draws on five authoritative clinical sources: Evidence-based guidelines for diagnosis and treatment of diabetic retinopathy in China (2022) [42], Guidelines for Laser Treatment of Fundus Diseases (Li & Liao) [43], Laser Treatment of Fundus Diseases (Zhang) [44], Guideline for the prevention and treatment of diabetes mellitus in China (2024 edition) [45], and Clinical Application Guidelines for National Essential Medicines: Chemical Drugs and Biological Products (2018 Edition) [46]. Through AI-assisted extraction and manual review, the contents of the guidelines are cleaned, formatted, and structured. Only treatment-related information is retained and subsequently divided into text chunks based on clinical topic similarity, thereby constructing specialized knowledge bases tailored for the four types of expert roles.

Retrieval Optimization To overcome the limitations of traditional vector retrieval in processing complex medical text, we adopt a summarization-based retrieval strategy. First, the agent distills a summary from the input text, which serves as a query to locate relevant knowledge segments. Then, the retrieved content undergoes a secondary refinement and reorganization by the LLM to filter out redundant noise. The refined knowledge, along with the original input, is fed into the core LLM. This dual-filtering mechanism ensures the professional relevance of the context provided to the primary model.

By incorporating a knowledge base, the single-agent architecture effectively mitigates model hallucinations and supports the clinical reliability of the proposed regimens.

4.2.3 Multi-Agent Collaboration

DRAgent simulates the treatment pathway for a real DR patient: The patient presents at the outpatient clinic with diagnostic information and fundus images. The outpatient physician compiles a clinical record and passes it to the ophthalmologist, who drafts an initial treatment plan and then consults with the laser surgeon and the endocrinologist in turn. The nurse integrates the outcomes of these two discussions into an updated plan with a consistent structure. The ophthalmologist further

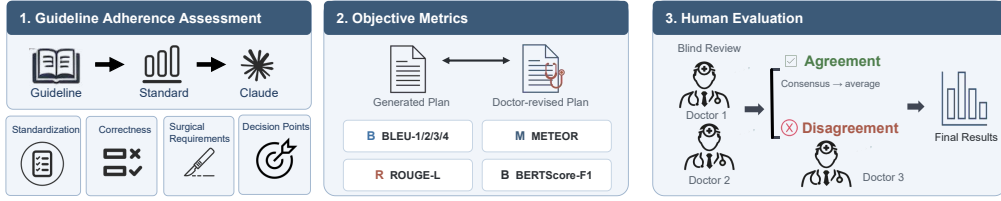


Fig. 5 The multidimensional evaluation framework

discusses medication usage and plan standardization with the pharmacist based on this updated version. The nurse integrates the feedback once more to generate the final treatment plan. The discussion process adheres to the following principles:

- The ophthalmologist consults the three specialists in sequence. In each round, both the ophthalmologist and the specialist must respond.
- A discussion ends when the consulted specialist fully agrees with the current plan, or when three rounds have elapsed.
- If the maximum number of rounds is reached without a consensus, the ophthalmologist’s plan is retained.

4.3 Multidimensional Evaluation Framework

To assess the quality of treatment plans generated by standalone LLMs and by DRAgent, we construct a multidimensional evaluation framework spanning guideline-compliance, objective metrics, and human evaluation, jointly balancing clinical normativity, textual accuracy, and practical feasibility (Fig. 5).

4.3.1 Guideline Compliance Evaluation

Clinical practice guidelines represent the standardized compliance reference for treatment planning. We extract core recommendations from five authoritative guidelines and, in line with the clinical characteristics of DR patients, design two complementary compliance schemes: one for ophthalmic treatment and one for endocrine management. The ophthalmology scheme covers four dimensions, namely normativity, correctness, surgery-specific requirements, and critical decision points, and the endocrinology scheme covers three dimensions, namely normativity, correctness, and follow-up monitoring. Within each dimension, multiple evaluation items probe specific aspects of plan quality, and each item is assigned a weight according to its clinical importance. Two experienced clinicians jointly review and confirm the full criteria, which are listed in Supplementary Tables 1 and 2.

A dynamic triggering mechanism activates evaluation items based on each patient’s pathological features, so that more complex cases trigger more items and a higher maximum score. Scoring is performed automatically by a third-party LLM (Claude Opus 4.8) that is independent of all models under test, following the predefined criteria and a dedicated scoring prompt (Supplementary Note 3).

$$\text{NCR} = \frac{\text{Actual Score}}{\text{Maximum Score of Triggered Items}} \times 100\%. \quad (1)$$

The NCR ranges from 0% to 100%; a higher NCR indicates closer adherence to guideline recommendations.

4.3.2 Objective Metrics

The objective evaluation dataset comprises two parts: the 100 multi-center real-world clinical cases, and 588 hybrid real-synthetic cases sampled from the 1,040-case hybrid subset. Although the hybrid cases originate from LLM-generated drafts, every treatment plan is revised by senior ophthalmologists

(associate chief physician or above) from Beijing Tongren Hospital; the revised plans are therefore adopted as reliable ground truth.

Taking each patient’s clinical record as input, we set the real clinical plans and the physician-revised plans as the reference standard and quantify plan quality by measuring the textual and semantic overlap between the generated plans and the references. We use standard implementations of BLEU-1 to BLEU-4 [47], ROUGE-L [48], METEOR [49] and BERTScore-F1 [50].

These metrics assume that a good answer closely resembles the reference—an assumption that often breaks down in medicine. Given the heterogeneity across cases and the variation in physicians’ writing styles, surface-level token matching cannot capture the substantive similarity between two plans. More critically, such metrics are insensitive to locally decisive errors: “perform surgical treatment” and “do not perform surgical treatment” share nearly all tokens and thus score highly, yet represent opposite clinical decisions. We therefore distill the core diagnostic and therapeutic information from each plan and compute similarity over the distilled content, suppressing interference from writing style and level of detail.

A complete plan comprises disease description, treatment principles, specific protocols, follow-up plan, patient education, and precaution. All are necessary, but their importance differs: patient education is largely similar across cases, whereas the treatment protocol and follow-up plan carry the clinical decisions and are the primary focus of evaluation. We thus adopt a weighted similarity scheme—similarity is computed per distilled item and aggregated by a weighted sum, with weights kept consistent with the expert-determined guideline-compliance weights above (Supplementary Table 3; distillation prompt in Supplementary Note 4). This ensures the applicability of semantic-similarity metrics to treatment-plan text, allowing them to effectively quantify the semantic agreement between the ophthalmic components of a generated plan and the ground-truth reference.

4.3.3 Human Evaluation

Human evaluation follows an independent dual-review protocol with third-expert adjudication. Three ophthalmologists of associate-senior rank or above, none of whom participated in dataset annotation, score plans blind through a custom-built clinical annotation system. For each patient, the two plans (one from a standalone LLM, one from DRAgent) are randomly assigned as Plan A or Plan B to remove order and preference effects, and two reviewers score them independently.

Clinical rationality is rated on a five-point scale: 1, entirely unreasonable, requiring complete rewriting; 2, partly unreasonable, requiring substantial revision; 3, basically reasonable, usable as a preliminary reference after optimization; 4, largely reasonable, requiring only minor revision; 5, highly reasonable, directly usable for clinical reference. Higher scores denote greater clinical rationality. Reviewers also indicate which plan is better ($A > B$, $B > A$, or $A = B$). When the two reviewers agree on the relative ranking, the final score is the mean of their ratings. When they disagree, a third, more senior ophthalmologist adjudicates independently, and the final score is the mean of the two concurring ratings among the three.

The three evaluation modes are complementary. Guideline-compliance assessment captures plan rationality at the level of overall structure and clinical norms; reference-based objective metrics offset the limited capacity of guidelines to represent individual variation, bringing evaluation closer to real-world practice; and human evaluation supplies the clinical judgement that automated metrics cannot, as such metrics tend to penalize linguistically divergent but equally reasonable plans. Human evaluation is the most reliable but also the most labor-intensive component; combining it with the automated measures yields an evaluation system that is both efficient and precise.

Ethics

This retrospective study was approved by the Medical Ethics Committee of Gaoyao District People’s Hospital, Zhaoqing, China (approval number: 202500; approved 8 April 2025). Ethical review of the Zhongshan Ophthalmic Center subset by the Institutional Review Board of Zhongshan Ophthalmic Center, Sun Yat-sen University, is currently in progress, and the approval number will be provided

prior to publication. Written informed consent was waived given the retrospective design and the use of fully de-identified data. All procedures adhered to the Declaration of Helsinki.

Data Availability

A subset of the DR treatment-planning dataset is available at <https://github.com/TiTiPlu/DRAgent>, comprising the complete hybrid real-synthetic subset, a sample of the de-identified real clinical cases, and a sample of the large-scale automatically generated cases. The complete dataset will be released upon acceptance of this manuscript. The fundus images themselves are not redistributed; they are available from their original repositories under the respective licenses (EyeQ/EyePACS, OIA-DDR, MMRDR, IDRiD and APTOS 2019). We release the case identifiers required to link our clinical records and treatment plans to the source images.

Code Availability

The code for DRAgent is publicly available at <https://github.com/TiTiPlu/DRAgent>.

References

- [1] Baxter, S. L., Gali, H. E., Huang, A. E. *et al.* Time requirements of paper-based clinical workflows and after-hours documentation in a multispecialty academic ophthalmology practice. *American Journal of Ophthalmology* **206**, 161–167 (2019).
- [2] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K. *et al.* Large language models in medicine. *Nature Medicine* **29**, 1930–1940 (2023).
- [3] Singhal, K., Azizi, S., Tu, T. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
- [4] Singhal, K., Tu, T., Gottweis, J. *et al.* Toward expert-level medical question answering with large language models. *Nature Medicine* **31**, 943–950 (2025).
- [5] Gong, E. J., Bang, C. S., Lee, J. J. & Baik, G. H. Knowledge-practice performance gap in clinical large language models: Systematic review of 39 benchmarks. *Journal of Medical Internet Research* **27**, e84120 (2025).
- [6] Dinc, M. T., Bardak, A. E., Bahar, F. & Noronha, C. Comparative analysis of large language models in clinical diagnosis: performance evaluation across common and complex medical cases. *JAMIA Open* **8**, ooaf055 (2025).
- [7] Rao, A. S., Esmail, K. P., Lee, R. S. *et al.* Large language model performance and clinical reasoning tasks. *JAMA Network Open* **9**, e264003 (2026).
- [8] Kim, Y., Park, C., Jeong, H. *et al.* Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems* **37**, 79410–79452 (2024).
- [9] Gao, S., Zhu, R., Kong, Z. *et al.* Txagent: An ai agent for therapeutic reasoning across a universe of tools. *arXiv preprint arXiv:2503.10970* (2025).
- [10] Wang, L. *et al.* A survey on large language model based autonomous agents. *Frontiers of Computer Science* **18**, 186345 (2024).
- [11] Qiu, J. *et al.* Llm-based agentic systems in medicine and healthcare. *Nature Machine Intelligence* **6**, 1418–1420 (2024).

- [12] Tang, X. *et al.* Medagents: Large language models as collaborators for zero-shot medical reasoning. *Findings of the Association for Computational Linguistics: ACL 2024* 599–621 (2024).
- [13] Chen, X., Yi, H., You, M. *et al.* Enhancing diagnostic capability with multi-agents conversational large language models. *npj Digital Medicine* **8**, 159 (2025).
- [14] Zhao, W. *et al.* An agentic system for rare disease diagnosis with traceable reasoning. *Nature* **651**, 775–784 (2026).
- [15] Kim, J., Lee, B., Sohn, K.-A., Kim, D. & Lee, Y. C. Evaluation of multi-agent llms in multidisciplinary team decision-making for challenging cancer cases. *Proceedings of the 10th Machine Learning for Healthcare Conference (PMLR)* **298** (2025). URL <https://proceedings.mlr.press/v298/kim25a.html>.
- [16] Li, J., Lai, Y., Li, W. *et al.* Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957* (2024).
- [17] Fan, Z. *et al.* Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)* 10183–10213 (2025).
- [18] Jin, D., Pan, E., Oufattole, N. *et al.* What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences* **11**, 6421 (2021).
- [19] Pal, A., Umapathi, L. K. & Sankarasubbu, M. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. *Proceedings of the Conference on Health, Inference, and Learning* 248–260 (2022).
- [20] Jin, Q., Dhingra, B., Liu, Z., Cohen, W. & Lu, X. PubMedQA: A dataset for biomedical research question answering. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* 2567–2577 (2019).
- [21] Qiu, P. *et al.* Quantifying the reasoning abilities of llms on clinical cases. *Nature Communications* **16**, 9799 (2025).
- [22] Wang, S., He, X., Jian, Z., Li, J. & Xu, C. Advances and prospects of multi-modal ophthalmic artificial intelligence based on deep learning: a review. *Eye and Vision* **11**, 38 (2024).
- [23] Shi, D. *et al.* A multimodal visual–language foundation model for computational ophthalmology. *npj Digital Medicine* **8**, 381 (2025).
- [24] Li, S. *et al.* Eyecaregpt: Boosting comprehensive ophthalmology understanding with tailored dataset, benchmark and model. *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)* 3893–3902 (2025).
- [25] Zhao, H., Ling, Q., Pan, Y. *et al.* Ophtha-llama2: A large language model for ophthalmology. *arXiv preprint arXiv:2312.04906* (2023).
- [26] Olawade, D. B., Weerasinghe, K., Mathugamage, M. D. D. E. *et al.* Enhancing ophthalmic diagnosis and treatment with artificial intelligence. *Medicina* **61**, 433 (2025).
- [27] Huang, L. *et al.* A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* **43**, 1–55 (2025).

- [28] Fu, H. *et al.* Evaluation of retinal image quality assessment networks in different color-spaces. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* **11764**, 48–56 (2019).
- [29] EyePACS & California Healthcare Foundation. Diabetic retinopathy detection. Kaggle (2015). URL <https://www.kaggle.com/c/diabetic-retinopathy-detection>. Fundus image dataset.
- [30] Li, T. *et al.* Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences* **501**, 511–522 (2019).
- [31] Tang, Z. *et al.* A multimodal retinal image dataset for diabetic retinopathy detection using foundation models. *Scientific Data* **13**, 639 (2026).
- [32] Porwal, P. *et al.* Indian diabetic retinopathy image dataset (idrid): A database for diabetic retinopathy screening research. *Data* **3**, 25 (2018).
- [33] Asia Pacific Tele-Ophthalmology Society. Aptos 2019 blindness detection. Kaggle (2019). URL <https://www.kaggle.com/competitions/aptos2019-blindness-detection>. Fundus image dataset.
- [34] Engelmann, J., Storkey, A. & Bernabeu, M. O. Quickqual: Lightweight, convenient retinal image quality scoring with off-the-shelf pretrained models. *Ophthalmic Medical Image Analysis (OMIA 2023)* **14096**, 32–41 (2023).
- [35] Singh, A. *et al.* Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267* (2025).
- [36] Comanici, G. *et al.* Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).
- [37] Liu, A. *et al.* Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).
- [38] Team, Q. Qwen3.5: Accelerating productivity with native multimodal agents (2026). URL <https://qwen.ai/blog?id=qwen3.5>.
- [39] FreedomIntelligence. HuatuoGPT-3: RL-only domain adaptation from base models via off-policy seeding. Hugging Face model card (2026). URL <https://huggingface.co/FreedomIntelligence/HuatuoGPT-3-32B>.
- [40] Anthropic. Claude opus 4.8. Model card (2026). Used in this study only as the third-party scorer for guideline-compliance evaluation.
- [41] Zakka, C., Shad, R., Chaurasia, A. *et al.* Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI* **1**, AIoa2300068 (2024).
- [42] Fundus Disease Group, Ophthalmology Branch, Chinese Medical Association & Fundus Disease Group, Ophthalmologist Branch, Chinese Medical Doctor Association. Evidence-based guidelines for diagnosis and treatment of diabetic retinopathy in china (2022). *Chinese Journal of Ocular Fundus Diseases* **39**, 99–124 (2023). In Chinese.
- [43] Li, X. & Liao, J. *Guidelines for Laser Treatment of Fundus Diseases* (People’s Medical Publishing House, Beijing, 2009). In Chinese.
- [44] Zhang, H. *Laser Treatment of Fundus Diseases* (People’s Medical Publishing House, Beijing, 2012). In Chinese.
- [45] Chinese Diabetes Society. Guideline for the prevention and treatment of diabetes mellitus in china (2024 edition). *Chinese Journal of Diabetes Mellitus* **17**, 16–139 (2025). In Chinese.

- [46] National Essential Medicines Clinical Application Guidelines and Formulary Editorial Board. *Clinical Application Guidelines for National Essential Medicines: Chemical Drugs and Biological Products (2018 Edition)* (People’s Medical Publishing House, Beijing, 2019). In Chinese.
- [47] Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* 311–318 (2002).
- [48] Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out* 74–81 (2004).
- [49] Banerjee, S. & Lavie, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* 65–72 (2005).
- [50] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

Author Contributions

T.S. conceived the study, designed the DRAgent framework and the multidimensional evaluation protocol, constructed the DR treatment-planning dataset, performed the experiments and statistical analyses, and wrote the original draft. Z.C., P.C. and L.L. contributed to the system implementation, the evaluation pipeline and the analysis of the results. Z.J., T.Z., Y.W. and H.X. collected and curated the multi-center real-world clinical dataset and participated in the clinical discussions that shaped the treatment-plan templates and the study design. M.T., Z.W. and J.Y. reviewed and revised the clinical records and treatment plans, defined the guideline-compliance criteria, organized and supervised the blinded human evaluation, and participated in the clinical discussions. W.X. and X.T. supervised the study, provided funding and resources, and revised the manuscript. All authors reviewed, revised and approved the final manuscript.

Acknowledgements

This study was supported by the National Key Research and Development Program of China (2023YFC2415400); the Major Key Projects of PCL under Grant (PCL2025A12-2); the National Natural Science Foundation of China (T2422012, 62071210); the Guangdong Basic and Applied Basic Research Foundation (2024B1515020088); the Shenzhen Science and Technology Program (RCYX20210609103056042); the Guangdong Basic and Applied Basic Research Foundation (2021A1515220131); and the High Level of Special Funds (G030230001, G03034K003).

Competing Interests

The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SUPPLEMENTARY.pdf](#)