npj | digital medicine

Perspective

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-026-02755-7

Rethinking scale in ophthalmic artificial intelligence: from bigger models to smarter clinical reasoning

Kai Jin[1,2] ✉, Kaikai Zhao[1,2,3], Rupesh Agrawal[4,5,6,7,8], Gui-Shuang Ying[9] & Andrzej Grzybowski[10,11]

Recent advances in ophthalmic AI have improved benchmark performance, yet clinical trust remains limited. We argue that progress should move beyond data and model scaling toward trustworthy, skill-efficient systems that integrate multimodal evidence, external knowledge, and uncertainty-aware reasoning. Ophthalmology provides a strong testbed for agentic AI, but safe clinical translation will require rigorous validation, workflow integration, and evaluation frameworks aligned with real-world decision making.

Over the past decade, advances in ophthalmic artificial intelligence (AI) have been largely driven by improvements in performance metrics such as sensitivity, specificity, accuracy, and area under the receiver operating characteristic curve (AUC)[1]. Expanding training datasets, increasing model capacity, and refining network architectures have led to steady gains across a wide range of ophthalmic imaging tasks, including diabetic retinopathy screening, glaucoma detection, and age-related macular degeneration classification[2]. These developments have firmly established AI as a powerful tool for visual pattern recognition in ophthalmology.

Despite growing regulatory approvals, such as FDA-cleared systems for diabetic retinopathy in the United States and over 20 CE-marked devices in Europe, real-world adoption lags markedly. This disparity highlights a "translation gap" in which benchmark success and regulatory clearance fail to overcome practical barriers such as workflow integration, reimbursement, and clinician trust[3]. Beyond these structural hurdles, as many ophthalmic AI systems approach performance saturation on curated benchmarks, it has become increasingly clear that numerical accuracy alone does not equate to clinical readiness. Models that achieve high benchmark scores may still behave inconsistently across devices, populations, or clinical contexts, and may fail in ways that are difficult for clinicians to anticipate or interpret[4,5]. As a result, the central question facing the field is no longer how to further improve performance, but how to build AI systems that clinicians can trust and are willing to use in routine clinical care.

This shift reflects a deeper re-evaluation of what "scale" means in ophthalmic AI. Progress can no longer be defined solely by dataset size or model performance, but increasingly by three complementary dimensions: trustworthiness, reasoning capability, and clinical skill efficiency. In this context, trustworthiness encompasses robustness under distribution shift, calibration of confidence, consistency across clinically similar cases, and safety in deployment. Reasoning capability reflects the ability to integrate multimodal evidence, incorporate external knowledge, and perform multi-step inference. Clinical skill efficiency captures how effectively an AI system transforms heterogeneous clinical information into actionable support for diagnosis, monitoring, or management. These dimensions can be evaluated through complementary domains: trustworthiness (robustness, calibration, consistency, and safety), reasoning capability (multimodal integration, retrieval fidelity, and multi-step coherence), and clinical skill efficiency (actionable decision support, longitudinal utility, and human-AI collaboration quality).

The limits of data-centric scaling

The prevailing paradigm in ophthalmic AI has been data-centric scaling, where larger datasets and increasingly complex models are expected to improve performance[6]. While this strategy has driven rapid progress, it is beginning to show diminishing returns. In many imaging benchmarks, adding more data yields only marginal improvements, suggesting the emergence of a ceiling effect. For several ophthalmic tasks, model performance has already approached the level of inter-grader agreement among expert clinicians, which represents a practical upper bound for many diagnostic benchmarks[7,8]. Consequently, further gains in

[1]Eye Center of Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China. [2]Zhejiang Provincial Key Laboratory of Ophthalmology, Zhejiang Provincial Clinical Research Center for Eye Diseases, Zhejiang Provincial Engineering Institute on Eye Diseases, Hangzhou, China. [3]School of Computer Science and Technology/School of Artificial Intelligence, China University of Mining and Technology, Xuzhou, China. [4]National Healthcare Group Eye Institute, Tan Tock Seng Hospital, Tan Tock Seng, Singapore, Singapore. [5]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore. [6]Singapore Eye Research Institute, Singapore, Singapore. [7]Duke-NUS Medical School, Singapore, Singapore. [8]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [9]Center for Preventive Ophthalmology and Biostatistics, Department of Ophthalmology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [10]Institute for Research in Ophthalmology, Foundation for Ophthalmology Development, Poznan, Poland. [11]Department of Ophthalmology, University of Warmia and Mazury, Olsztyn, Poland. ✉e-mail: jinkai@zju.edu.cn

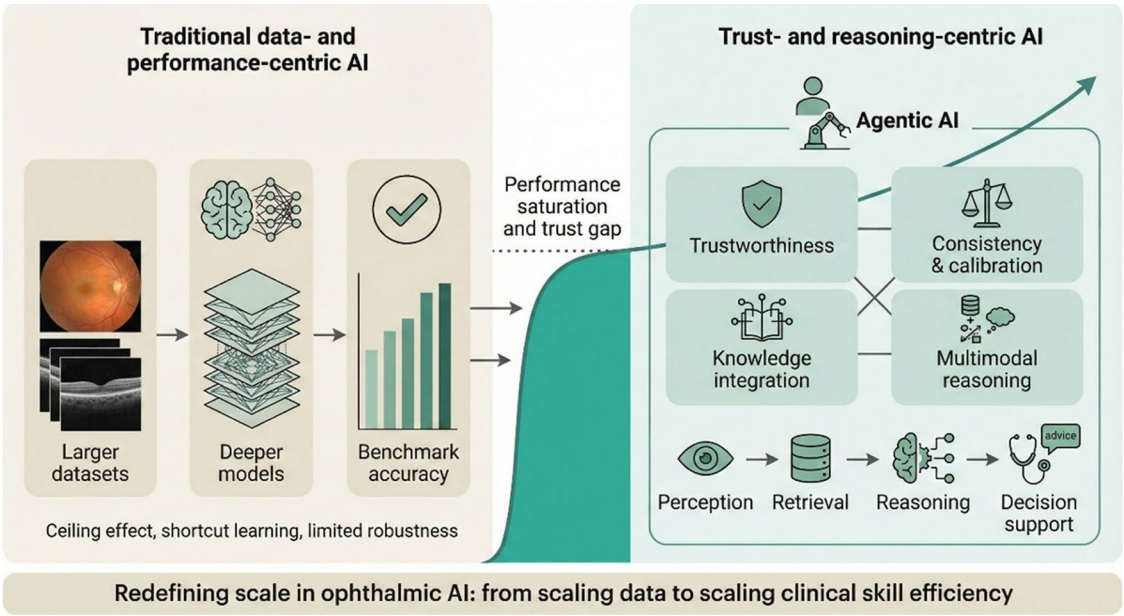

Fig. 1 | Conceptual illustration of a paradigm shift in ophthalmic artificial intelligence. Traditional performance-centric scaling emphasizes larger datasets and deeper models to optimize benchmark accuracy, but increasingly encounters performance saturation, shortcut learning, and limited robustness. In contrast, a trust- and reasoning-centric paradigm prioritizes trustworthiness, consistency, knowledge integration, and agentic reasoning, enabling artificial intelligence systems to transform information into clinical skills and support decision-making under human oversight. This shift reframes scale from data and model size toward clinical skill efficiency. This figure was created by the authors using CorelDRAW.

numerical accuracy may translate into minimally into clinically meaningful improvements.

A fundamental limitation lies in how deep learning models learn. These systems optimize statistical correlations rather than causal understanding, and frequently exploit shortcuts that are predictive within a given dataset but lack true clinical relevance[9]. Such shortcut learning can inflate benchmark performance while obscuring vulnerability to real-world variation (Fig. 1).

In ophthalmology, where imaging devices, acquisition protocols, and disease presentations vary widely, these limitations become particularly evident. A model trained on millions of images may still struggle when confronted with an unfamiliar scanner, a rare phenotype, or a subtle longitudinal change. Models trained on curated public datasets may also underperform when applied to heterogeneous real-world clinical data, as demonstrated in glaucoma classification studies assessing generalizability across public and routine-care datasets[10]. Similarly, multicenter real-world validation studies of diabetic retinopathy screening systems have revealed substantial variation in performance and gradability across algorithms and deployment settings[11]. These findings suggest that scaling data alone does not guarantee robust clinical intelligence. Recent ophthalmic electronic health record studies further suggest that progress need not rely solely on increasing the size of labeled datasets. Clinically informed semi-supervised learning has shown strong performance across varying labeled-data fractions, highlighting that meaningful gains may also come from more effective use of unlabeled clinical data rather than from simply expanding labeled datasets[12]. Although improving robustness across devices, populations, and clinical environments may still require large and diverse datasets for validation, the key shift is that scale should increasingly reflect diversity, representativeness, efficient use of unlabeled real-world data, and evaluation rigor rather than simply the raw number of labeled training samples.

Trustworthiness as a new axis of progress

As performance gains plateau, trustworthiness has emerged as a more clinically relevant dimension of progress. Trustworthiness extends beyond accuracy to include robustness under distribution shift, calibration, consistency across similar cases, interpretability, and alignment with clinical reasoning[13].

From a clinical perspective, not all errors are equivalent. False negatives that miss vision-threatening disease may carry substantially greater clinical consequences than minor disagreements between adjacent severity classes. Evaluation strategies that incorporate asymmetric scoring rules, risk-weighted metrics, or uncertainty estimation may therefore provide a more clinically meaningful assessment of model behavior. An AI system that behaves predictably, expresses uncertainty appropriately, and supports human decision-making may therefore be more valuable than one that achieves marginally higher accuracy but fails unpredictably.

These challenges the field to move beyond leaderboard-driven development. Rather than asking which model performs best on a benchmark, the more meaningful question is whether the system behaves in a manner that clinicians can understand, anticipate, and safely integrate into practice.

From pattern recognition to clinical skill efficiency

True clinical intelligence is defined not by the volume of data a system has consumed, but by how efficiently it transforms information into clinical skills. Human ophthalmologists do not rely solely on memorized patterns; they reason, integrate multimodal evidence, search for relevant knowledge, and adapt to uncertainty.

Most current ophthalmic AI systems, including large foundation models, remain primarily pattern recognizers. Even when trained on diverse datasets, their strengths lie in representation learning rather than explicit reasoning. As a result, they may achieve strong performance without demonstrating an understanding of disease context, progression, or management principles.

Reconceptualizing progress in terms of clinical skill efficiency, the ability to generalize knowledge, reason across modalities, and adapt to unfamiliar scenarios, offers a more clinically meaningful notion of scale (Fig. 1)[14]. This perspective shifts emphasis from learning more patterns to learning how to reason.

Toward agentic AI in ophthalmology

Agentic AI represents a transition from passive prediction to active clinical reasoning[15,16]. Rather than producing isolated outputs, agentic systems can autonomously retrieve relevant information, integrate external medical

Table 1 | Contrasting paradigms in ophthalmic artificial intelligence

Dimension	Performance-centric AI	Trust- and reasoning-centric AI
Primary objective	Maximize benchmark performance	Enable trustworthy clinical decision support
Definition of scale	Larger datasets and deeper models	Reasoning capability and clinical skill efficiency
Learning focus	Statistical correlation and pattern recognition	Knowledge integration and clinical reasoning
Generalization	Highly dataset-dependent	Context-aware and task-adaptive
Robustness	Vulnerable to domain and device shift	Adaptive to uncertainty and variability
Model behavior	Static inference	Agentic, multi-step reasoning
Clinical role	Automated prediction tool	Decision-support partner under human oversight
Failure characteristics	Confident but opaque errors	Explicit uncertainty and transparent reasoning
Readiness for unknown cases	Limited	Improved adaptability through reasoning

knowledge, perform multi-step reasoning, and revise conclusions as new evidence emerges[17]. Early implementations in ophthalmology, such as EyeAgent, a multimodal LLM-based agent that orchestrates specialized tools for comprehensive clinical decision support across diverse imaging modalities, demonstrate these capabilities in practice[18].

However, the current evidence base remains limited, and several challenges must be addressed before widespread clinical deployment. These include hallucination risks in large language models, failures in tool selection or knowledge retrieval, latency and cost, medico-legal liability, and the difficulty of validating agentic systems within clinical longitudinal workflows. Therefore, agentic AI should be viewed as a promising but still immature paradigm rather than an established solution.

In ophthalmology, such capabilities could substantially expand the clinical role of AI[19]. An agentic system might analyze imaging data, retrieve clinical guidelines, compare longitudinal findings, and reason about disease progression or management strategies[20]. Importantly, this approach offers a mechanism for handling uncertainty and novelty, enabling AI systems to respond more safely to cases that fall outside their training distribution.

Agentic AI does not imply unsupervised decision-making. Rather, it provides a framework in which AI systems can support clinicians through transparent reasoning processes, while remaining embedded within human oversight and clinical governance. Table 1 contrasts the key characteristics of the traditional performance-centric paradigm with the proposed trust- and reasoning-centric approach.

Importantly, these systems are best viewed as adaptive clinical copilots rather than autonomous decision-makers, with their value lying in structured reasoning support rather than the replacement of clinician judgment.

Why ophthalmology is a critical testbed

Ophthalmology presents a particularly compelling environment for reasoning-driven AI[21]. Clinical decisions frequently require the integration of multiple imaging modalities, such as color fundus photography, optical coherence tomography, angiography, and visual fields, alongside longitudinal follow-up and contextual clinical information.

Diseases such as glaucoma and age-related macular degeneration illustrate the limitations of static, image-centric models. Diagnosis and management depend not only on isolated images, but on progression patterns, risk stratification, and individualized treatment goals. AI systems optimized solely for single-image performance are unlikely to meet these clinical demands.

By contrast, agentic and reasoning-oriented AI systems align more closely with the realities of ophthalmic practice, where uncertainty, heterogeneity, and evolving disease trajectories are the norm rather than the exception.

Implications for evaluation and regulation

This shift in perspective has important implications for how ophthalmic AI should be evaluated and regulated. Accuracy on benchmark datasets, while necessary, is insufficient. Evaluation frameworks should increasingly incorporate measures of robustness, calibration, reasoning consistency, and safety under distribution shift[22]. Such paradigms may also need to include clinically weighted error analysis, uncertainty quantification, longitudinal consistency, and human-AI interaction studies that assess decision-support quality rather than prediction accuracy alone.

Regulatory approaches may also need to evolve to accommodate adaptive and reasoning-enabled systems. Ensuring transparency, traceability of reasoning processes, and appropriate human oversight will be essential for responsible deployment in clinical settings. Such evaluation paradigms may require moving beyond static test sets toward stress testing under domain shift, longitudinal consistency analysis, and human-AI interaction studies that assess decision support quality rather than prediction accuracy alone.

Conclusion

As ophthalmic AI matures, progress should no longer be defined primarily by the scale of data or model performance. The next phase of advancement lies in scaling trust, reasoning capability, and clinical skill efficiency. Moving beyond performance-driven development toward agentic, knowledge-integrated, reasoning-driven systems may ultimately determine whether AI becomes a reliable clinical partner rather than a high-scoring but fragile tool.

Data availability

No datasets were generated or analysed during the current study.

Received: 10 February 2026; Accepted: 4 May 2026;
Published online: 10 May 2026

References

1. Jin, K. & Ye, J. Artificial intelligence and deep learning in ophthalmology: current status and future perspectives. *Adv. Ophthalmol. Pract. Res.* **2**, 100078 (2022).
2. Li, Z. et al. Artificial intelligence in ophthalmology: the path to the real-world clinic. *Cell Rep. Med.* **4**, 101095 (2023).
3. Grauslund, J. & Grzybowski, A. The seven sins of automated diabetic retinopathy screening: barriers for clinical implementation of artificial intelligence-based devices in national screening programs. *Ophthalmologica* **248**, 137–140 (2025).
4. Tan, T. F. et al. Artificial intelligence and digital health in global eye health: opportunities and challenges. *Lancet Glob. Health* **11**, e1432–e1443 (2023).
5. Wenderott, K., Krups, J., Zaruchas, F. & Weigl, M. Effects of artificial intelligence implementation on efficiency in medical imaging—a systematic literature review and meta-analysis. *npj Digit. Med.* **7**, 265 (2024).
6. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).

7. Wang, T.-W. et al. Systematic review and meta-analysis of regulator-approved deep learning systems for fundus diabetic retinopathy detection. *npj Digit. Med.* **9**, 110 (2025).
8. Srinivasan, S. et al. Inter-observer agreement in grading severity of diabetic retinopathy in wide-field fundus photographs. *Eye* **37**, 1231–1235 (2023).
9. Grzybowski, A., Jin, K. & Wu, H. Challenges of artificial intelligence in medicine and dermatology. *Clin. Dermatol.* **42**, 210–215 (2024).
10. Rashidisabet, H. et al. Validating the generalizability of ophthalmic artificial intelligence models on real-world clinical data. *Transl. Vis. Sci. Technol.* **12**, 8 (2023).
11. Moradi, M. et al. Clinically informed semi-supervised learning improves disease annotation and equity from electronic health records: a glaucoma case study. *npj Digit. Med.* **9**, 82 (2025).
12. Lee, A. Y. et al. Multicenter, head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. *Diabetes Care* **44**, 1168–1175 (2021).
13. González-Gonzalo, C. et al. Trustworthy AI: closing the gap between development and integration of AI systems in ophthalmic practice. *Prog. Retin. Eye Res.* **90**, 101034 (2022).
14. Srinivasan, S. et al. Ophthalmological question answering and reasoning using OpenAI o1 versus other large language models. *JAMA Ophthalmol* **143**, 740–748 (2025).
15. Gao, S. et al. Empowering biomedical discovery with AI agents. *Cell* **187**, 6125–6151 (2024).
16. Abou Ali, M., Dornaika, F. & Charafeddine, J. Agentic AI: a comprehensive survey of architectures, applications, and future directions. *Artif. Intell. Rev.* **59**, 11 (2025).
17. Zou, J. & Topol, E. J. The rise of agentic AI teammates in medicine. *Lancet* **405**, 457 (2025).
18. Chen, X. et al. EyeAgent: a large language model-based AI agent for multimodal multitask ophthalmology analyses. *Invest. Ophthalmol. Vis. Sci.* **66**, 4641 (2025).
19. Liu, F. et al. A foundational architecture for AI agents in healthcare. *Cell Rep. Med.* **6**, 102374 (2025).
20. Qiu, Y. et al. Embodied artificial intelligence in ophthalmology. *npj Digit. Med.* **8**, 351 (2025).
21. Xu, P. et al. DeepSeek-R1 outperforms Gemini 2.0 Pro, OpenAI o1, and o3-mini in bilingual complex ophthalmology reasoning. *Adv. Ophthalmol. Pract. Res.* **5**, 189–195 (2025).
22. Lim, Z. W. et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine* **95**, 104770 (2023).

Acknowledgements

This study was supported by National Natural Science Foundation of China (82201195).

Author contributions

Kai Jin conceived the idea and led the writing of the manuscript. Kaikai Zhao conducted the literature analysis and prepared the figures. Rupesh Agrawal, Gui-Shuang Ying, and Andrzej Grzybowski contributed through critical review and editing. All authors approved the final manuscript.

Competing interests

Kai Jin and Rupesh Agrawal are Associate Editors of *npj Digital Medicine* and were not involved in the selection of peer reviewers for this manuscript nor in any subsequent editorial decisions. The other authors do not have a competing interest.

Additional information

Correspondence and requests for materials should be addressed to Kai Jin.