AMERICAN ACADEMY OF OPHTHALMOLOGY®

In-Context Learning for Data-Efficient Diabetic Retinopathy Detection via Multimodal Foundation Models

Murat S. Ayhan, PhD,[1,2] *Ariel Y. Ong, MD,*[1,2] *Eden Ruffell,*[1,2] *Siegfried K. Wagner, MD, PhD,*[1,2] *David A. Merle, MD,*[1,2,‡] *Pearse A. Keane, MD*[1,2,‡,]

Objective: This study aims to evaluate whether in-context learning (ICL), a prompt-based learning mechanism enabling multimodal foundation models to rapidly adapt to new tasks without retraining or large annotated datasets, can achieve comparable diagnostic performance to domain-specific foundation models. Specifically, we use diabetic retinopathy (DR) detection as an exemplar task to probe if a multimodal foundation model (Google Gemini 1.5 Pro), employing ICL, can match the performance of a domain-specific model (RETFound) fine-tuned explicitly for DR detection from color fundus photographs (CFPs).

Design: A cross-sectional study.

Subjects: A retrospective, publicly available dataset (Indian Diabetic Retinopathy Image Dataset) comprising 516 CFPs collected at an eye clinic in India, featuring both healthy individuals and patients with DR.

Methods: The images were dichotomized into 2 groups based on the presence or absence of any signs of DR. RETFound was fine-tuned for this binary classification task, while Gemini 1.5 Pro was assessed for it under zero-shot and few-shot prompting scenarios, with the latter incorporating random or k-nearest-neighbors-based sampling of a varying number of example images. For experiments, data were partitioned into training, validation, and test sets in a stratified manner, with the process repeated for 10-fold cross-validation.

Main Outcome Measures: Performance was assessed via accuracy, F1 score, and expected calibration error of predictive probabilities. Statistical significance was evaluated using Wilcoxon tests.

Results: The best ICL performance with Gemini 1.5 Pro yielded an average accuracy of 0.841 (95% confidence interval [CI]: 0.803–0.879), an F1 score of 0.876 (95% CI: 0.844–0.909), and a calibration error of 0.129 (95% CI: 0.107–0.152). RETFound achieved an average accuracy of 0.849 (95% CI: 0.813–0.885), an F1 score of 0.883 (95% CI: 0.852–0.915), and a calibration error of 0.081 (95% CI: 0.066–0.097). While accuracy and F1 scores were comparable ($P > 0.3$), RETFound's calibration was superior ($P = 0.004$).

Conclusions: Gemini 1.5 Pro with ICL demonstrated performance comparable to RETFound for binary DR detection, illustrating how future medical artificial intelligence systems may build upon such frontier models rather than being bespoke solutions.

Financial Disclosure(s): Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2026;6:100934 © 2025 by the American Academy of Ophthalmology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).*

Supplemental material available at www.ophthalmologyscience.org.

Foundation models trained on broad, multimodal data have demonstrated impressive performance in diverse scenarios[1–5] and hold promise as general-purpose solutions adaptable to various clinical tasks.[6] However, their deployment in specialized domains like medical imaging often requires resource-intensive fine-tuning on domain-specific datasets, limiting accessibility and scalability in all settings, not just in resource-constrained ones.[7]

Initially observed in natural language processing[8] and later in computer vision,[9–12] in-context learning (ICL) offers an alternative to transfer learning that typically requires fine-tuning of a pretrained model to new tasks before making predictions. Unlike transfer learning via fine-tuning, ICL does not require model retraining or parameter updates and instead allows foundation models to be adapted to new tasks by conditioning them via task-specific prompts.[13] These prompts typically include concise task descriptions along with a few illustrative examples.[14]

In-context learning is particularly relevant for medical use cases, as language-based prompting allows nontechnical users to easily adjust a model's behavior. By eliminating the need for extensive labeled datasets, computational resources, or coding expertise, ICL can expand access to cutting-edge artificial intelligence (AI), allowing clinicians and researchers to leverage powerful models developed outside health care without the burden of

https://doi.org/10.1016/j.xops.2025.100934
ISSN 2666-9145/25

building custom solutions from scratch. Recent reports of widespread adoption of the DeepSeek models[15] in Chinese hospitals exemplify this trend.[16] Additionally, ICL can potentially enhance explainability by exploiting multimodal foundation models' ability to generate natural language-based descriptions of clinically relevant features and decision rationales, aligning AI outputs with the interpretability needs of clinicians and stakeholders.

Recently, ICL has been shown to enable classification of pathology images while attaining performance comparable to or surpassing that of fine-tuned foundation models, despite using significantly fewer annotated examples.[17] Given that ophthalmology is an imaging-driven specialty of medicine, we explore the potential of ICL in this field by using it to detect signs of diabetic retinopathy (DR) on color fundus photographs (CFPs) in a proof-of-concept study. Diabetic retinopathy detection was selected not due to unmet diagnostic performance, but because it serves as a well-suited benchmark for evaluating new learning approaches.[18,19]

We demonstrate that ICL using a multimodal foundation model (Gemini 1.5 Pro) can achieve diagnostic performance comparable to RETFound, a domain-specific foundation model specialized for retinal imaging,[1] while also providing fairly well-calibrated predictive uncertainty estimates simply via prompt engineering. In addition, we provide evidence that the Gemini model can offer a window into its decision mechanism through counterfactual reasoning, achieved solely through prompt engineering.

Methods

Dataset

We evaluated the performance of ICL against transfer learning using a well-known publicly available dataset, the Indian Diabetic Retinopathy Image Dataset (IDRiD).[20] Five hundred sixteen macula-centered images were acquired in mydriasis via a Kowa VX-10α digital fundus camera with a 50° field of view and were taken from clinical examinations performed at an eye clinic in India.[21] All images in the dataset were graded according to the International Clinical Diabetic Retinopathy Severity Scale[22] by 2 medical experts who provided adjudicated consensus grades.[21] No demographic information, for example, age, sex, or ethnicity, was available in the public dataset.

DR Detection as a Binary Classification Task

To test ICL as a potential tool for DR detection, we defined our task as a binary classification problem by dichotomizing the severity labels into the following groups: {0} vs. {1,2,3,4}. Thus, in the presence of any signs of DR, a classifier is expected to assign the positive class label 1 (DR present). Otherwise, it should assign 0, the negative class label indicating an absence of DR. For probabilistic classification, this can be achieved by estimating $p(y = 1|x)$, where x is an image and the model, $f(x)$, essentially outputs the probability of the image belonging to the positive class, that is, $y = 1$. Then, a simple thresholding scheme yields the most likely class label: If $p(y = 1|x) \geq 0.5$, then 1; else 0.

A well-calibrated classifier provides probability estimates that accurately reflect the true likelihood of its predictions being correct. This ensures that its outputs can be interpreted as confidence values, making it easier to assess the reliability of automated decisions. Such classifiers can be integrated into clinical workflows,[23] aid in decision referrals, and highlight cases where clinical decision-making may be particularly challenging.[24–26] Expected calibration error (ECE) summarizes the overall calibration quality of a classifier into a single metric by capturing the gap between its confidence and accuracy.[27]

Model Development

We induced classifiers to perform the binary DR detection task via 2 approaches: transfer learning and ICL (Fig 1). Our code, including prompts tailored to the task, is also available at https://github.com/msayhan/ICL-Ophthalmology-Public.

Transfer Learning. For transfer learning, we used RETFound,[1] which was pretrained firstly on ImageNet[28] and then on 904 170 CFPs. We coupled its feature extraction encoder with a single linear layer for binary classification and fine-tuned the entire model using the binary cross-entropy loss in an end-to-end fashion, including all layers. In order to evaluate the model on IDRiD, we performed 10-fold stratified cross-validation. Specifically, we allocated 10% of data for testing, while another 10% went for validation and the remaining 80% for training. We trained each model for 20 epochs using binary cross-entropy loss, evaluating on the validation set after each epoch. We then selected the models with the smallest validation loss for testing. During training, we used a weight decay parameter of 0.01, an initial learning rate of 0.001, which is linearly scaled with the minibatch size of 8 times batch accumulation steps of 2 divided by 256, and a layer decay of 0.75. Our optimizer was AdamW[29] coupled with cosine scheduling and warm restarts in every 100 steps.[30] For data augmentation, we used standard transformations, including random cropping, brightness, contrast, saturation, and hue adjustments, as well as Gaussian blur and rotation. Lastly, we normalized pixel values via the ImageNet statistics.

ICL. All ICL experiments in this study were performed with Google's Gemini model (Gemini 1.5 Pro). Given the temperature range [0, 2.0] for this particular model and its default value of 1.0, we trialed several values, including 0.1, 0.3, 0.5, 0.6, 0.7, and 0.75, and settled on 0.7 in order to slightly trade the model's randomness off against its determinism in responses. Additionally, we adopted nucleus sampling (also known as top-p sampling)[31] with a threshold of 0.9 for the probability mass of most likely tokens to be generated. The remaining and potentially unreliable portions of probability distributions were truncated in the hope of avoiding degenerate text.[31]

We used the stratified partitions described earlier (80% for training, 10% for validation, and 10% for testing, repeated 10 times) also for the ICL experiments. Considering individual images from test sets as query objects, say x, we simply prompted the model to classify them one at a time according to the absence or presence of DR pathology in zero-shot settings. For few-shot learning, we additionally sampled support sets from the nontest partitions with k examples, where $k \epsilon \{3, 5, 10, 20\}$, from each class (negative and positive) and concatenated them with prompts. Sampling was either random or based on k-nearest-neighbors (kNNs) of a query image. When using kNN, we used fundus image feature representations extracted via RETFound and sorted images in ascending order with respect to their cosine distance to the query image. The top k images were returned and used as examples in prompts. For supervision in few-shot learning, we also padded images as follows during concatenation: "Ophthalmologists classified the following image as {y}: {x}," where y was the class label as either "normal" or "diabetic retinopathy (DR)."

We used three main prompts: system instruction (also known as master prompt), zero-shot prompt, and few-shot prompt (see

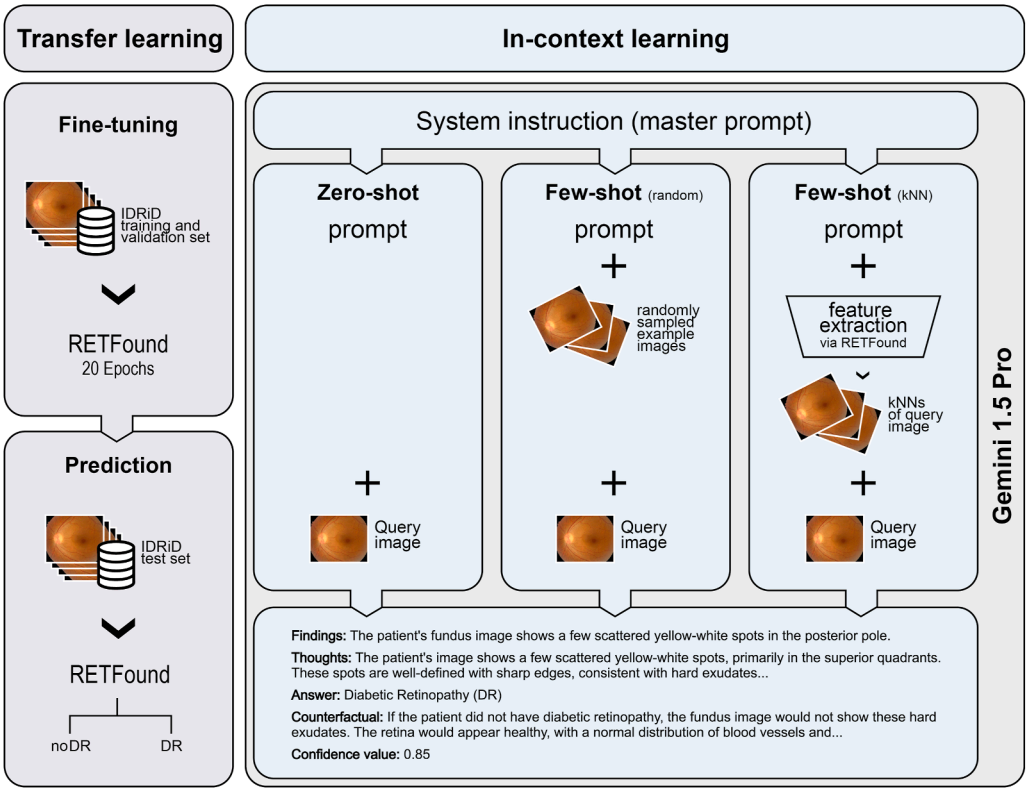

Figure 1. Study workflow. We performed binary DR classification using the IDRiD dataset. Transfer learning with RETFound served as a comparator (left panel). We used ICL with Gemini 1.5 Pro (right panel), employing 3 distinct prompting strategies: zero-shot prompting, few-shot prompting with randomly selected example images, and few-shot prompting with kNN-based example selection. Across all ICL scenarios, a system instruction (master prompt) was provided to define the model's role. For details on dataset splitting and training procedures, refer to the methodological explanations provided in the Methods section. DR = diabetic retinopathy; ICL = in-context learning; IDRiD = Indian Diabetic Retinopathy Image Dataset; kNN = k-nearest-neighbor.

Supplementary File 1 available at www.ophthalmologyscience.org). The system instruction was used when a Gemini model was instantiated, and it described a role for the model to assume, for example, "a helpful and professional medical assistant for an ophthalmologist who needs to classify CFPs of patients." The system instruction also included examples of good and bad responses as well as the overall structure of expected JavaScript Object Notation (JSON) outputs. During zero-shot prompting, the prompt elaborated on the role and defined the classification task along with domain-specific considerations regarding retinal structures, DR pathology, and its appearance on CFPs, which were then followed by concrete steps for the analysis of images and decision-making. Finally, we reiterated the description of JSON output format by specifying details on the fields like "findings," "thoughts," "answer," and "counterfactual," as well as "confidence_value" for the answer given at that instance. For few-shot prompting, we tried to steer the model to utilize additional images by including additional instructions such as "carefully examine examples and find patterns that distinguish normal images from diseased ones" and "compare what you see in the patient's image to the patterns you learned from the examples," while keeping most of the prompt identical to the zero-shot one. For counterfactual reasoning, we encouraged the model in both zero-shot and few-shot scenarios to ponder alternative scenarios through questions like "If the patient had not had DR, how would the image have looked?" or "If the patient had had DR, how would the image have looked?"

Results

We evaluated the performance of both RETFound and Gemini 1.5 Pro on test sets via accuracy, F1 score, and the ECE of predictive probabilities (Fig 2). RETFound achieved an average accuracy of 0.849 (95% confidence interval [CI]: 0.813−0.885), an average F1 score of 0.883 (95% CI: 0.852−0.915), and an average ECE of 0.081 (95% CI: 0.066−0.097). For Gemini 1.5 Pro, we started out with rudimentary prompts (Supplementary File 1A) that included only broad, high-level descriptions of retinal structures, DR pathology, and general CFP features. While the rudimentary prompts (Fig 2, green lines) resulted in an average accuracy of 0.641 (95% CI: 0.610−0.673), an average F1 score of 0.547 (95% CI: 0.496−0.599), and an average ECE of 0.348 (95% CI: 0.340−0.357) in the zero-shot (k = 0) setting, the performance increased with few-shot prompting and reached an average accuracy of 0.773 (95% CI: 0.739−0.807), an F1 score of 0.738 (95% CI: 0.695−0.780), and an average ECE of 0.287 (95% CI: 0.274−0.300) with kNN-based sampling of 20 examples per class. There was no significant difference in performance between random or kNN-based sampling strategies across different values of k.

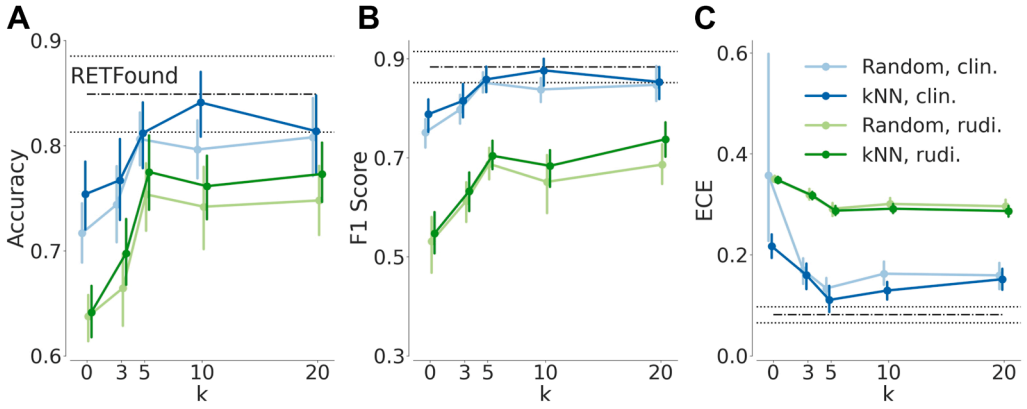

Figure 2. Diabetic retinopathy classification performance of ICL in comparison with RETFound. Mean performances based on 10-fold cross-validation are shown along with 95% confidence intervals. Horizontal black lines (dash-dotted for the mean and dotted for the confidence interval) indicate RETFound's performance. Categorical plots in blue or green show the ICL results with varying numbers of examples. Blue indicates results achieved with optimized prompts, whereas green indicates results obtained by using rudimentary (rudi.) prompts. (**A**) Accuracy, (**B**) F1 score, and (**C**) ECE via relplot.[36] ECE = expected calibration error; ICL = in-context learning; kNN = k-nearest-neighbor.

Next, we iteratively optimized prompts by incorporating increasingly detailed clinical descriptions of general features visible on CFPs along with descriptions of specific DR-associated alterations. In addition, we provided a clear framework for classification (Supplementary File 1B). The optimized prompts dramatically improved the ICL performance of Gemini 1.5 Pro and resulted in an average accuracy of 0.754 (95% CI: 0.713−0.794), an average F1 score of 0.788 (95% CI: 0.748−0.829), and an average ECE of 0.217 (95% CI: 0.188−0.246) in the zero-shot (k = 0) setting (Fig 2, blue lines). With few-shot prompting and kNN-based sampling of 10 images per class, the model's ICL performance peaked at an average accuracy of 0.841 (95% CI: 0.803−0.879), an average F1 score of 0.877 (95% CI: 0.844−0.909), and an average ECE of 0.129 (95% CI: 0.107−0.152). There was no significant difference in the classification performance of RETFound and Gemini 1.5 Pro guided with clinical knowledge and few relevant examples (*P* values for accuracy and F1 score: 0.326 and 0.432, respectively). RETFound's predictive probabilities were, however, significantly better calibrated (*P* value: 0.004).

Despite optimized prompts, ICL with few-shot prompting with random image sampling was almost never competitive with transfer learning via RETFound. For the F1-score, ICL only reached the RETFound performance level with k = 20 examples (*P* value: 0.106). In terms of accuracy, RETFound was better (*P* value: 0.049). In contrast, kNN-based sampling with k = 5 led to an ICL calibration performance noninferior to that of RETFound (*P* value: 0.106).

We used both models' predictions from test runs and computed their confusion matrices for the whole collection of 516 CFPs (Fig 3). On these predictions, RETFound's sensitivity and specificity were 0.862 and 0.821, respectively. Gemini 1.5 Pro achieved 0.845 and 0.833 for the same measures. With a Cohen kappa score of 0.700, the agreement between models was substantial.[32]

Finally, we selected examples to demonstrate the outputs of Gemini 1.5 Pro with ICL and kNN-based sampling (Fig 4). The examples in the first 2 rows belong to the "DR" and "no DR" (normal) classes, respectively. The model's textual descriptions of retinal findings and its thoughts and counterfactual reasoning provide insights into how the model arrived at correct decisions on these 2 cases. Its answer for the last example, however, was wrong. Interestingly, it was able to assign the correct label with random sampling.

Discussion

We demonstrated that Gemini 1.5 Pro, a general-purpose multimodal foundation model, achieved performance comparable to RETFound, a domain-specific model for retinal imaging that was fine-tuned with a linear head on a small dataset, on a binary DR detection task. This was accomplished using an ICL approach, with as few as 10 representative CFPs from each class providing sufficient context to guide inferences. These findings underscore the untapped potential of multimodal foundation models for enabling timely translation of cutting-edge AI developments into clinical research by side-stepping the need for *de novo* model development or fine-tuning as well as cumbersome efforts of data annotation at scale. Notably, this study has not even leveraged the most advanced multimodal large language models available today, yet has still achieved results comparable to a domain-specific model. As foundation models continue to evolve, their applications in medical imaging could extend far beyond current expectations.

During the experiments, we realized that adding more examples sometimes led to a paradoxical dip in performance. Despite being nonsignificant, this could indicate that focusing on a judiciously selected set of examples may help the model better concentrate on key features for diagnostic decision-making. On that note, selection of examples can also impact the model's performance beyond implicit assumptions. Recent work in visual ICL has emphasized the importance of prompt structure and example selection in determining model performance. Studies have shown that curated or retrieval-

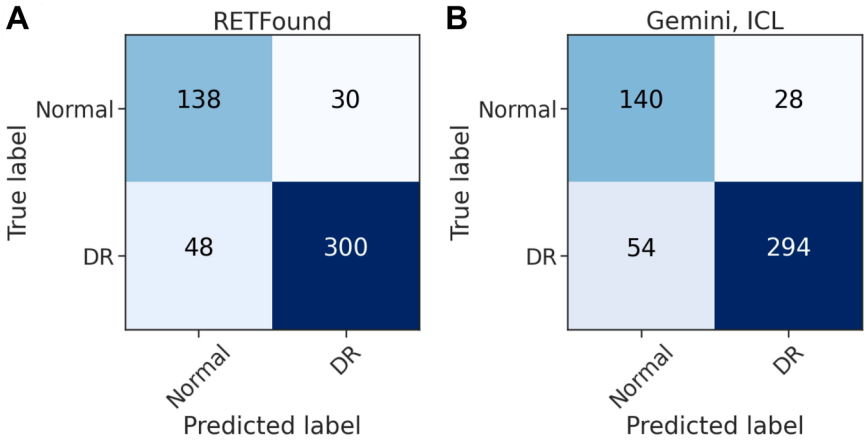

Figure 3. Confusion matrices for RETFound and Gemini-ICL predictions on the IDRiD dataset during test time. (**A**) Confusion matrix for RETFound. (**B**) Confusion matrix for Gemini 1.5 Pro using ICL. In-context learning was performed with k = 10 NNs per class. DR = diabetic retinopathy; ICL = in-context learning; IDRiD = Indian Diabetic Retinopathy Image Dataset; kNN = k-nearest-neighbor.

based examples consistently outperform random sampling and that sophisticated prompting strategies can improve generalization across vision tasks.[33–35] However, selection of examples can also impact the model's performance beyond implicit assumptions. For instance, for the DR example in Fig 4C, kNN-based sampling resulted in predominantly normal-looking images, despite the presence of DR, thereby potentially biasing the model's decision toward the incorrect classification (Supplementary File 2 and Supplementary Figure 1, available at www.ophthalmologyscience.org). This effect is particularly relevant when applying kNN sampling to images that contain clinically significant pathologies, such as neovascularization or intraretinal microvascular abnormalities, but otherwise exhibit a relatively mild appearance (e.g., only a few hemorrhages or microaneurysms). In such cases, the strongly pathological features may be overshadowed by the overall normal-like appearance, causing kNN-based sampling to select images that resemble healthy cases (Supplementary File 2 and Supplementary Figure 1). On the other hand, random sampling returned more prominent examples of DR, and the model made the correct decision based on them (Supplementary File 2 and Supplementary Figure 2, available at www.ophthalmologyscience.org). Thus, future studies may benefit from a hybrid approach where both random and kNN-based examples are used together for ICL. In scenarios where performance is highly sensitive to example selection, a supervised retrieval framework[35] can also be used to determine the examples most aligned with the ICL purposes and mitigate the impact of suboptimal examples.

We also explored how prompting the model to re-engage with the same image through counterfactual reasoning led to slight yet consistent improvements in classification accuracy. Although these gains were not large, the approach mirrors human diagnostic workflows, where specialists frequently revisit initial assessments to verify or refine their conclusions. This iterative style of interrogation could eventually serve as a means to mitigate oversights, thereby increasing both clinician confidence and patient safety. Moreover, Gemini 1.5 Pro's ability to generate written explanations for its decisions represents an additional advantage, as it offers clinicians or educators a language-based and therefore easily accessible window into the reasoning process. Such text-based justifications can highlight the visual cues the model deems most relevant, facilitating human review of the model's decision-making steps and potentially accelerating the education of trainees through interactive case discussions. However, future experiments will have to critically assess how reliable those explanations are.

On a more general level, clinical deployment of such tools faces notable challenges, including privacy concerns, limited local compute resources, and stringent regulatory requirements, particularly for large language model-based systems. While widespread clinical adoption may not yet be feasible, the text-based interaction paradigm offers immediate potential to support clinician-led research and exploratory use.

Looking ahead, the growing capabilities of general-purpose foundation models raise the possibility of supporting a broad range of clinical and research tasks without the need to develop and fine-tune separate models for each application. Our findings illustrate that, with appropriate prompting and minimal task-specific examples, such models can approximate the performance of specialized tools for a defined task. While significant challenges remain for clinical implementation, this approach may enable more flexible and scalable use of AI, particularly in research settings where adaptability and rapid prototyping are valuable.

Limitations

Despite the intriguing results, this study has several limitations. The proprietary nature of Gemini 1.5 Pro's training data makes it unclear whether prior exposure to the IDRiD dataset influenced results. This also leads to transparency concerns, as the model's training corpus remains undisclosed, limiting interpretability and bias assessment. Therefore, we cannot categorically exclude the possibility that prior familiarity with the IDRiD dataset during

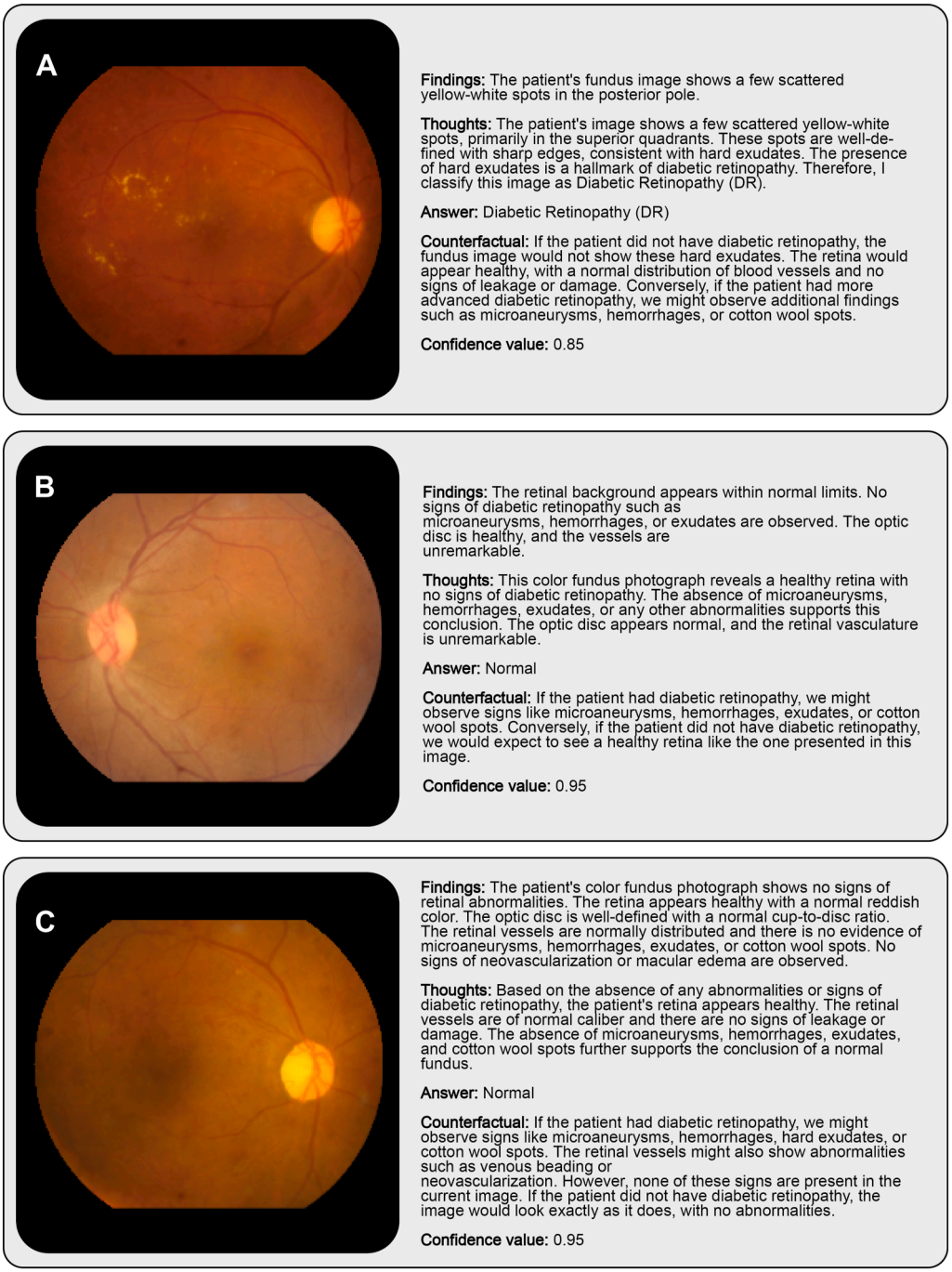

Figure 4. Representative example images and Gemini-ICL (k = 10 NNs per class) outputs. (**A**) A case of correctly identified DR. (**B**) A case of correctly identified healthy fundus appearance. (**C**) A DR image misclassified as healthy when using kNN-based sampling. DR = diabetic retinopathy; ICL = in-context learning; kNN = k-nearest-neighbor.

pretraining may have influenced model performance. However, the conspicuous gap in performance between rudimentary prompts (see Fig 2 and Supplementary File 1) and carefully engineered prompts implies that skilled prompt design itself, rather than mere exposure to the images, plays a central role in eliciting the model's diagnostic capabilities. Another important observation is that the model struggled with fine-grained detection, precise enumeration, or meticulous observation of small pathological features, a limitation that may pose challenges in advanced DR staging, which can hinge on identifying and counting subtle lesions. The binary classification task is a relatively simple scenario, and it remains uncertain how well ICL would perform in more complex multiclass settings. In addition, the study relied on a single, modestly sized dataset (IDRiD), and external validation was not

performed. This limits the assessment of generalizability across different populations, imaging conditions, and disease distributions. Future research should also explore its performance in multitask scenarios, where models are expected to solve different tasks simultaneously,[24] validate findings in diverse datasets as well as clinical settings, and assess newer model iterations. Newer members of the Gemini family, for example, the now available Gemini 2.5 Pro or other state-of-the-art models, along with better adaptation strategies, may well lead to improved speed, accuracy, and reasoning depth, as well as enhanced capabilities for numerical tasks. Another consideration for limitation is that increasing the number of support examples did not always improve classification and sometimes reduced accuracy, highlighting the need for careful selection to mitigate bias, particularly with kNN-based sampling. While accuracy and F1 scores were comparable to RETFound, Gemini 1.5 Pro exhibited poorer calibration, which may impact the reliability of its confidence scores in clinical applications. Lastly, it would have been optimal to tune the temperature and prompt configurations on a validation set, with final evaluation restricted to the test set. Consequently, our reported results may be marginally overestimated owing to leakage from the test set. Nevertheless, given that ICL is applied postmodel training, such that the underlying model weights remain unaltered, and the test partition is internal to the IDRiD dataset, we maintain confidence in the validity of our qualitative conclusions. Additionally, running multiple trials per instance could better capture variability in responses and improve performance estimates but was not feasible due to computational constraints. Future work should incorporate such variability and adopt more sophisticated evaluation protocols for clinical validation.

Conclusion

This study highlights the potential of ICL with multimodal foundation models for medical AI, demonstrating that Gemini 1.5 Pro can match the performance of a domain-specific foundation model in DR classification without retraining or large annotated datasets. Its ability to generate language-based explanations enhances interpretability and educational value. While challenges remain in transparency, calibration, and fine-grained pathology detection, these findings suggest that multimodal foundation models could transform medical imaging by enabling scalable, data-efficient diagnostic support across diverse medical applications.

Acknowledgments

The authors acknowledge support through UKRI EPSRC (Artificial intelligence innovation to accelerate health research, EP/Y017803/1 [M.S.A.]), National Institute for Health Research (NIHR)—Moorfields Eye Charity (MEC) Doctoral Fellowship (NIHR303691 [A.Y.O.]), University College London UKRI Centre for Doctoral Training in AI-enabled healthcare systems Studentship (EP/S021612/1 [E.R.]), EURETINA (Retinal Medicine Clinical Research Grant [D.A.M. and P.A.K.]), UK Research & Innovation Future Leaders Fellowship (MR/T019050/1 [P.A.K.]), and The Rubin Foundation Charitable Trust (P.A.K.).

Footnotes and Disclosures

Originally received: March 20, 2025.
Final revision: August 22, 2025.
Accepted: August 26, 2025.
Available online: September 3, 2025. Manuscript no. XOPS-D-25-00182.

[1] Moorfields Eye Hospital NHS Foundation Trust, NIHR Moorfields Biomedical Research Centre, London, UK.

[2] Institute of Ophthalmology, University College London, London, UK.

[‡]D.A.M. and P.A.K. contributed equally and share last authorship.

Disclosures:

All authors have completed and submitted the ICMJE disclosures form.

The authors made the following disclosures:

P.A.K.: Consultant − Retina Consultants of America, Topcon, Roche, Boehringer-Ingelheim; Payment or honoraria − Zeiss, Topcon, Novartis, Boehringer-Ingelheim, Apellis, Roche, AbbVie; Travel expenses − Bayer, Topcon, Roche; Participation on a Data Safety Monitoring Board − Topcon, Bayer, Boehringer-Ingelheim, RetinAI, Novartis; Stock/stock options − Big Picture Medical, Bitfount.

 Support was provided through UKRI EPSRC (Artificial intelligence innovation to accelerate health research, EP/Y017803/1 [M.S.A.]), National Institute for Health Research (NIHR)—Moorfields Eye Charity (MEC) Doctoral Fellowship (NIHR303691 [A.Y.O.]), University College London UKRI Centre for Doctoral Training in AI-enabled healthcare systems Studentship (EP/S021612/1 [E.R.]), EURETINA (Retinal Medicine Clinical Research Grant [D.A.M. and P.A.K.]), UK Research & Innovation Future Leaders Fellowship (MR/T019050/1 [P.A.K.]), and The Rubin Foundation Charitable Trust (P.A.K.).

Data Availability: The data set analyzed in this study is publicly available from https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid. All prompts used in the analyses are published alongside this manuscript.

Support for Open Access publication was provided by University College London.

HUMAN SUBJECTS: Human subjects were included in this study. This study used only publicly available, deidentified datasets. No new data were collected, and no identifiable information about individuals was accessed or processed. As such, they did not require institutional review board or ethics committee approval. Informed consent was not applicable. The research was conducted in accordance with the principles of the Declaration of Helsinki.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Ayhan, Merle

Data collection: Ayhan, Ong, Ruffell, Wagner, Merle, Keane

Analysis and interpretation: Ayhan, Ong, Ruffell, Wagner, Merle, Keane

Obtained funding: N/A

Overall responsibility: Ayhan, Ong, Ruffell, Wagner, Merle, Keane

Abbreviations and Acronyms:
AI = artificial intelligence; **CFP** = color fundus photograph; **CI** = confidence interval; **DR** = diabetic retinopathy; **ECE** = expected

calibration error; **ICL** = in-context learning; **IDRiD** = Indian Diabetic Retinopathy Image Dataset; **kNN** = k-nearest-neighbor.

Keywords:
AI, Artificial intelligence, Diabetic retinopathy, Large language model, LLM.

Correspondence:
Pearse A. Keane, MD, NIHR Biomedical Research Centre, Moorfields Eye Hosptial NHS Trust and University College London Institute of Ophthalmology, London, UK. E-mail: p.keane@ucl.ac.uk.

References

1. Zhou Y, Chia MA, Wagner SK, et al. A foundation model for generalizable disease detection from retinal images. *Nature*. 2023;622:156−163.
2. Chen RJ, Ding T, Lu MY, et al. Towards a general-purpose foundation model for computational pathology. *Nat Med*. 2024;30:850−862.
3. Lu MY, Chen B, Williamson DFK, et al. A multimodal generative AI copilot for human pathology. *Nature*. 2024;634:466−473.
4. Kondepudi A, Pekmezci M, Hou X, et al. Foundation models for fast, label-free detection of glioma infiltration. *Nature*. 2025;637:439−445.
5. Fallahpour A, Ma J, Munim A, et al. *MedRAX: Medical Reasoning Agent for Chest X-ray*. arXiv; 2025.
6. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. *Nature*. 2023;616:259−265.
7. Sevgi M, Ruffell E, Antaki F, et al. Foundation models in ophthalmology: opportunities and challenges. *Curr Opin Ophthalmol*. 2025;36:90−98.
8. Brown TB, Mann B, Ryder N, et al. *Language Models are Few-Shot Learners*. arXiv; 2020.
9. Shtedritski A, Rupprecht C, Vedaldi A. *What does CLIP know about a red circle? Visual prompt engineering for VLMs*. arXiv; 2023.
10. Ravi N, Gabeur V, Hu YT, et al. *SAM 2: Segment Anything in Images and Videos*. arXiv; 2024.
11. Zhang Y, Zhou K, Liu Z. What makes good examples for visual in-context learning?. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23*. NeurIPS Foundation; 2023:17773−17794.
12. Foster T, Croitoru I, Dorfman R, Edlund C, Varsavsky T, Almazán J. *Flexible Visual Prompts for In-Context Learning in Computer Vision*. arXiv; 2023.
13. Bommasani R, Hudson DA, Adeli E, et al. *On the Opportunities and Risks of Foundation Models*. arXiv; 2022.
14. Lampinen AK, Chan SCY, Singh AK, Shanahan M. *The broader spectrum of in-context learning*. arXiv; 2024.
15. DeepSeek AI, Guo D, Yang D, et al. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. arXiv; 2025.
16. Olcott E, Ding W. *DeepSeek spreads across China with Beijing's backing*. Financial Times Ltd; 2025.
17. Ferber D, Wölflein G, Wiest IC, et al. In-context learning enables multimodal large language models to classify cancer pathology images. *Nat Commun*. 2024;15:10104.
18. Ting DSW, Cheung CYL, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318:2211−2223.
19. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402−2410.
20. Porwal P, Pachade S, Kamble R, et al. Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. *Data*. 2018;3:25.
21. Porwal P, Pachade S, Kokare M, et al. IDRiD: diabetic retinopathy − segmentation and grading challenge. *Med Image Anal*. 2020;59:101561.
22. Wilkinson CP, Ferris FL, Klein RE, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*. 2003;110:1677−1682.
23. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. *J Med Ethics*. 2020;46:205−211.
24. Ayhan MS, Faber H, Kühlewein L, et al. Multitask learning for activity detection in neovascular age-related macular degeneration. *Transl Vis Sci Techol*. 2023;12:12.
25. Ayhan MS, Kühlewein L, Aliyeva G, et al. Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. *Med Image Anal*. 2020;64:101724.
26. Ayhan MS, Neubauer J, Uzel MM, et al. Interpretable detection of epiretinal membrane from optical coherence tomography with deep neural networks. *Sci Rep*. 2024;14:8484.
27. Guo C, Pleiss G, Sun Y, Weinberger KQ. *On Calibration of Modern Neural Networks*. arXiv; 2017.
28. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis*. 2015;115:211−252.
29. Loshchilov I, Hutter F. *Decoupled Weight Decay Regularization*. arXiv; 2019.
30. Loshchilov I, Hutter F. *SGDR: Stochastic Gradient Descent with Warm Restarts*. arXiv; 2017.
31. Holtzman A, Buys J, Du L, Forbes M, Choi Y. *The Curious Case of Neural Text Degeneration*. arXiv; 2020.
32. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159−174.
33. Shtedritski A, Rupprecht C, Vedaldi A. *What does CLIP know about a red circle? Visual prompt engineering for VLMs*. arXiv; 2023.
34. Foster T, Croitoru I, Dorfman R, Edlund C, Varsavsky T, Almazán J. *Flexible visual prompts for in-context learning in computer vision*. arXiv; 2023.
35. Zhang Y, Zhou K, Liu Z. *What Makes Good Examples for Visual In-Context Learning?*. arXiv; 2023.
36. Błasiok J, Nakkiran P. *Smooth ECE: Principled Reliability Diagrams via Kernel Smoothing*. arXiv; 2023.