RetSTA: An LLM-Based Approach for Standardizing Clinical Fundus Image Reports

Jiushen Cai[1], Weihang Zhang[1,3]*, Hanruo Liu[3,4,5], Ningli Wang[2,3,4], and Huiqi Li[1,3]

[1] Beijing Institute of Technology, Beijing, 100081, China
[2] Henan Academy of Innovations in Medical Science, Zhengzhou, 450000, China
[3] Beijing Key Laboratory of Intelligent Diagnosis Technology and Equipment for Optic Nerve-Related Eye Diseases, Beijing, 100069, China
[4] Beijing Tongren Hospital, Capital Medical University, Beijing, 100730, China
[5] Henan Eye Hospital, Henan Provincial People's Hospital, Zhengzhou, 450000, China
zhangweihang@bit.edu.cn

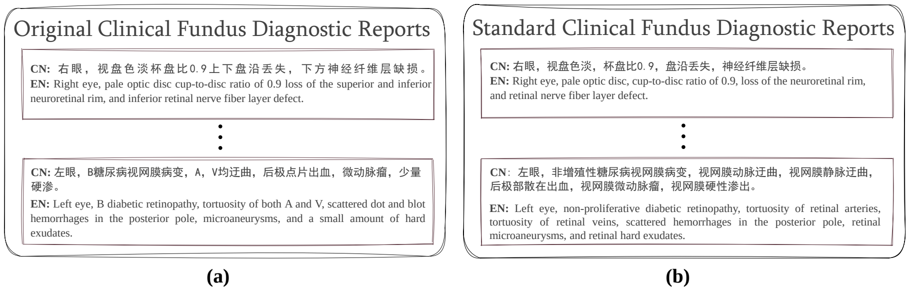
Abstract. Standardization of clinical reports is crucial for improving the quality of healthcare and facilitating data integration. The lack of unified standards, including format, terminology, and style, is a great challenge in clinical fundus diagnostic reports, which increases the difficulty for large language models (LLMs) to understand the data. To address this, we construct a bilingual standard terminology, containing fundus clinical terms and commonly used descriptions in clinical diagnosis. Then, we establish two models, RetSTA-7B-Zero and RetSTA-7B. RetSTA-7B-Zero, fine-tuned on an augmented dataset simulating clinical scenarios, demonstrates powerful standardization behaviors. However, it encounters a challenge of limitation to cover a wider range of diseases. To further enhance standardization performance, we build RetSTA-7B, which integrates a substantial amount of standardized data generated by RetSTA-7B-Zero along with corresponding English data, covering diverse complex clinical scenarios and achieving report-level standardization for the first time. Experimental results demonstrate that RetSTA-7B outperforms other compared LLMs in bilingual standardization task, which validates its superior performance and generalizability. The checkpoints are available at https://github.com/AB-Story/RetSTA-7B.

Keywords: Clinical Fundus Diagnostic Report · Large Language Model · Standardization.

1 Introduction

Clinical diagnostic reports are crucial records in medical practice, where their accuracy and clarity play a decisive role in patient diagnosis and treatment. Standardized clinical reports can reduce communication costs among healthcare professionals, minimize the potential for misunderstandings, and ensure medical

* Corresponding author



Original Clinical Fundus Diagnostic Reports

CN: 右眼，视盘色淡杯盘比0.9上下盘沿丢失，下方神经纤维层缺损。
EN: Right eye, pale optic disc cup-to-disc ratio of 0.9 loss of the superior and inferior neuroretinal rim, and inferior retinal nerve fiber layer defect.

⋮

CN: 左眼，B糖尿病视网膜病变，A，V均迂曲，后极点片出血，微动脉瘤，少量硬渗。
EN: Left eye, B diabetic retinopathy, tortuosity of both A and V, scattered dot and blot hemorrhages in the posterior pole, microaneurysms, and a small amount of hard exudates.

(a)

Standard Clinical Fundus Diagnostic Reports

CN: 右眼，视盘色淡，杯盘比0.9，盘沿丢失，神经纤维层缺损。
EN: Right eye, pale optic disc, cup-to-disc ratio of 0.9, loss of the neuroretinal rim, and retinal nerve fiber layer defect.

⋮

CN：左眼，非增殖性糖尿病视网膜病变，视网膜动脉迂曲，视网膜静脉迂曲，后极部散在出血，视网膜微动脉瘤，视网膜硬性渗出。
EN: Left eye, non-proliferative diabetic retinopathy, tortuosity of retinal arteries, tortuosity of retinal veins, scattered hemorrhages in the posterior pole, retinal microaneurysms, and retinal hard exudates.

(b)

Fig. 1. The importance of standardization: **(a)** Non-standardized reports lead to difficulties in comprehension. **(b)** Standardized reports enhance clarity and accuracy of understanding.

quality and patient safety [22]. They can be used to generate structured electronic medical records, providing data support for clinical research and labeled data for the development of artificial intelligence-based algorithms. Furthermore, standardized reports can facilitate the integration and analysis of medical data across multiple institutions, providing high-quality input for data mining and LLM development [12,13,8,21]. Modern ophthalmology research requires large volumes of high-quality data increasingly. It not only expands the available sample size for complete data analysis, but also significantly improves the quality, reliability, and efficiency of ophthalmic research endeavors [25]. As shown in Figure 1, the current lack of unified standards in the preparation of clinical diagnostic reports leads to significant variations in format, writing style, and terminology usage, which complicates the understanding of clinical data [14,16,3].

In clinical diagnostic process for fundus conditions, existing ophthalmic clinical guidelines and terminology fail to cover all scenarios encountered. Detailed descriptions of fundus disease symptoms heavily rely on the interpreting experience of ophthalmologists. Furthermore, the prevalence of negative and ambiguous expressions, typos, and issues related to system storage and formatting errors in reports exacerbate the challenges in standardizing fundus diagnostic reports.

Fortunately, the recent emergence of LLMs has introduced new research perspectives to this issue [26,31]. LLMs such as ChatGPT [2] and LLaMA [24] have demonstrated significant potential across various domains, leveraging their robust capabilities in natural language understanding and generation. DeepSeek has successively released multiple models, including DeepSeek-v3 [19], which rivals the performance of GPT-4o [1] at an exceptionally low cost. In medical field, LLMs like UltraMedical [30] and OpenBioLLM [20] have been utilized for tasks such as diagnosis, clinical report generation, and medical Q&A. LLMs have also demonstrated significant potential in the field of ophthalmology, offering promising prospects for advancing both ophthalmic research and clinical practice [23,7,6,5,4,18,17]. These advancements indicate that LLMs hold vast application prospects in the standardization of clinical diagnostic reports, addressing challenges arising from inconsistent report formats and terminology effectively.

To take the advantages of LLMs, we propose a novel approach to standardize fundus diagnostic reports. Our main contributions are as follows: Firstly, we construct a bilingual standard terminology for fundus clinical practice. This terminology integrates diverse clinical guidelines, standardized terms, and real diagnostic reports to ensure comprehensiveness and practicality. Secondly, we propose RetSTA-7B-Zero, a LLM specifically designed to comprehend complex expressions in fundus clinical diagnostic scenarios. This model demonstrates exceptional performance in handling intricate standardization tasks. Lastly, we propose RetSTA-7B, which is fine-tuned on 60,037 bilingual 'original report-standard report' pairs, ensuring its domain adaptability and disease coverage. To the best of our knowledge, this represents the first attempt to develop a standardization model at the report level, not even limited to ophthalmology.

2 Method

2.1 Standard Terminology

As shown in Figure 2, the fundus diagnostic terminology constructed in this paper comprehensively references international authoritative medical standards and real clinical diagnostic reports. Based on the International Classification of Diseases – 11 (ICD-11), Preferred Practice Pattern (PPP) issued by the American Academy of Ophthalmology, and the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT), the core terminology for fundus diseases was systematically integrated, encompassing fundamental aspects such as disease names, clinical symptoms, and critical examination indicators. By referencing these authoritative standards, the terminology ensures normative and international compatibility. Given that descriptions of certain symptoms in clinical practice are characterized by detail and dynamism, which standard terminology often fails to capture fully, we extracted frequently occurring descriptive phrases from clauses of 456,956 real fundus diagnostic reports using Affinity Propagation clustering[10]. These terms were preliminarily screened for accuracy, clarity, and clinical relevance, serving as a critical supplement to the standardized terminology.

To ensure the practicality and reliability of the terminology, we conducted rule-based matching and manual review of candidate terms under the guidance of a senior ophthalmologist. This process aims to eliminate duplicates or ambiguous expressions, and potentially misleading terms, preventing the coexistence of multiple terms for the same clinical entity and retaining only those with clear semantics. As a result, we successfully constructed a bilingual fundus standard clinical terminology, comprising 362 standardized terms.

2.2 Construction of RetSTA-7B-Zero

Fine-tuning LLMs for standardizing fundus clinical reports requires a substantial amount of labeled data, while in practical research, acquiring large-scale task-

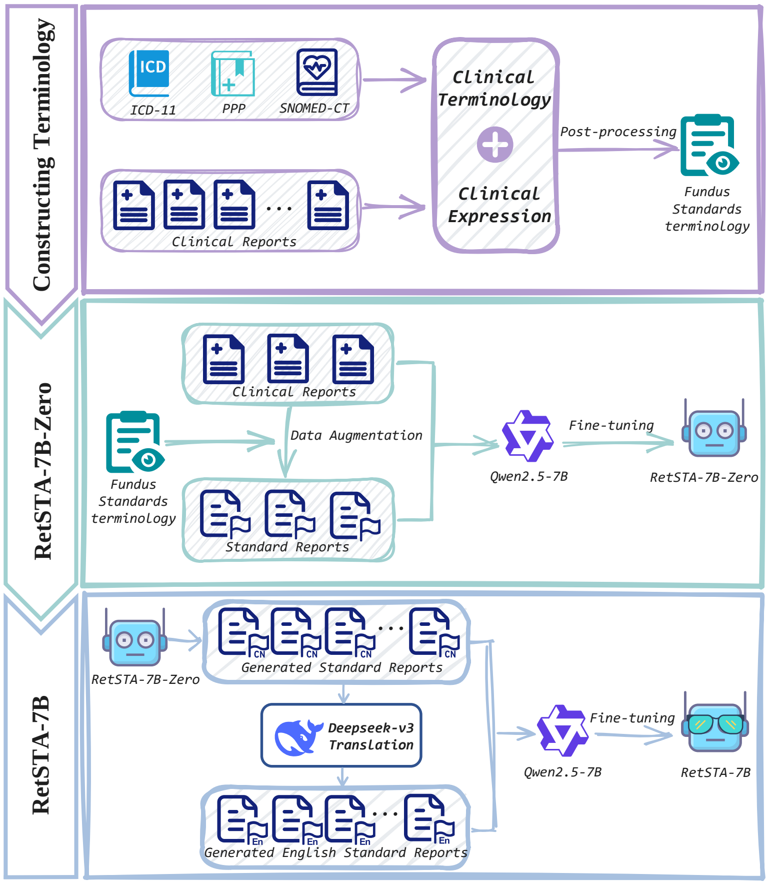


Fig. 2. Overview of the diagnostic reports standardization paradigm

relevant labeled data presents significant challenges. To alleviate this issue, we constructed 4,000 pairs of ’original report-standard report’ pairs, with 2,000 pairs serving as an independent test set to evaluate the model’s generalization performance, and the remaining 2,000 pairs forming the foundational training set.

To further expand the data scale, we designed a simple yet effective data augmentation strategy that accounts for various forms of noise prevalent in clinical scenarios, including typos, missing punctuation, ambiguous expressions, and negations. Building upon the foundational training set, this strategy employed techniques such as syntactic noise injection, semantic perturbation, and synonym replacement to expand the data volume to 10,000. Syntactic noise injection simulates real typos by inserting typos randomly, deleting punctuation, or splitting long sentences within the text. Semantic perturbation enhances the model’s ability to understand complex semantics by adding negation words or probabilistic modifiers. Synonym replacement utilizes a thesaurus of similar words to substitute disease descriptions, further increasing data diversity. We used the Low-Rank Adaptation (LoRA) [15] to fine-tune the Qwen2.5-7B-Instruct [28] model on the augmented dataset, resulting in the preliminary standardized model, RetSTA-7B-Zero.

2.3 RetSTA-7B for Standardization

To further enhance the model's coverage of clinical long-tail scenarios, we utilized RetSTA-7B-Zero to perform batch inference on 60,000 original fundus diagnostic reports, generating candidate standardized reports. To further improve the quality of the generated reports, we employed a rule-based filtering approach first. Specifically, generated reports were split into clauses by commas. Clauses were scored pre/post-standardization; those above a threshold were retained, otherwise deleted. Then, we calculated text similarity on the filtered reports and removed redundant data with description homogeneity above a threshold.

To enhance the model's cross-lingual adaptability and further improve its capability in medical scenarios, DeepSeek-V3 was employed to translate the Chinese dataset into an English version. During the translation process, we additionally implemented terminology consistency checks to ensure the translation accuracy, thereby mitigating semantic distortion caused by linguistic differences effectively. Ultimately, a mixed dataset containing 60,037 pairs of bilingual reports was constructed. This dataset not only covers standardized descriptions of various fundus diseases but also possesses cross-lingual compatibility. Based on the above dataset, we fine-tuned the Qwen2.5-7B-Instruct model using LoRA, ultimately developing RetSTA-7B, a LLM specifically tailored for the standardization of clinical fundus diagnostic reports.

3 Experiment

3.1 Experimental Setup

Datasets Two datasets were used for model fine-tuning in the experiments: a dataset of 10,000 samples for fine-tuning RetSTA-7B-Zero, and a bilingual dataset of 60,037 samples for fine-tuning RetSTA-7B. Additionally, an independent test set of 2,000 samples was used to evaluate model performance. For the evaluation of other LLMs, a unified 5-shot learning strategy was adopted, with the standard terminology provided as a reference during testing to ensure consistency and fairness.

Evaluation Metrics To comprehensively evaluate the performance of the model in the task of standardizing clinical fundus diagnostic reports, this paper employed commonly used metrics in the field of Natural Language Generation (NLG), including BLEU-1, BLEU-4, ROUGE-L, and METEOR.

Implementation Details We selected the 7B version of the pre-trained Qwen-2.5 model, obtaining its checkpoints directly from the official Huggingface repository. All models were trained on NVIDIA Geforce RTX 4090 GPU. The RetSTA-7B-Zero model was fine-tuned for 1 epoch, while the RetSTA-7B model was fine-tuned for 3 epochs. During fine-tuning, the batch size was set to 2, and the gradient accumulation steps were set to 8. We used the AdamW optimizer with 0.03 warm-up ratio and a learning rate of 3e-5.

Table 1. Results on English clinical diagnostic reports.

	BLEU-1	**BLEU-4**	**METEOR**	**ROUGE-L**
<10B Large Language Models				
UltraMedical-8B [30]	26.20	22.01	35.33	25.09
OpenBioLLM-8B [20]	61.82	47.17	52.01	44.02
Yi-1.5-9B-Chat [29]	70.47	56.86	59.29	52.58
GLM-4-9B [11]	76.40	62.18	59.76	57.46
Baichuan2-7B-Chat [27]	75.49	59.79	60.57	52.55
Llama-3.1-8B-Instruct [9]	53.01	43.29	53.27	43.85
Qwen2.5-7B-Instruct [28]	20.23	16.10	18.77	16.03
RetSTA-7B	**92.69**	**89.66**	**86.04**	**90.08**
>10B Large Language Models				
Qwen2.5-72B-Instruct [28]	80.32	66.65	66.69	57.27
GLM-4-Plus [11]	70.44	54.53	53.46	52.52
DeepSeek-V3 [19]	84.39	76.12	73.69	74.55

Baselines We compared our models with two types of LLMs: 1) Small LLMs: UltraMedical-8B [30], OpenBioLLM-8B [20], Yi-1.5-9B [29], GLM-4-9B [11], LLaMA-3.1-8B [9], Qwen-2.5-7B-Instruct [28], Baichuan2-7B [27]; and 2) Large LLMs: Qwen2.5-72B [28], GLM-4-Plus [11] and DeepSeek-V3 [19], where UltraMedical-8B [30] and OpenBioLLM-8B [20] are medical-specific LLMs.

3.2 Experimental Results

RetSTA-7B significantly outperforms the other nine compared LLMs on the test set, including two medical-specific LLMs and three LLMs with large-scale, highlighting its robust capability in the standardization task of clinical diagnostic report.

For English reports, the results are shown in Table 1. The results indicate that medical-specific LLMs, like UltraMedical [30] and OpenBioLLM [20], specialize in medical Q&A but struggle on the standardization task, even when the task is medically related. This suggests that standardization task of clinical diagnostic reports requires not only medical knowledge but also a deep understanding

Table 2. Results on Chinese clinical diagnostic reports.

	BLEU-1	**BLEU-4**	**METEOR**	**ROUGE-L**
<10B Large Language Models				
Yi-1.5-9B-Chat [29]	70.31	56.45	75.76	71.37
GLM-4-9B [11]	74.77	60.23	77.60	75.17
Llama-3.1-8B-Chinese-Chat [9]	79.02	66.11	80.21	78.88
Qwen2.5-7B-Instruct [28]	76.83	65.09	79.81	77.99
Baichuan2-7B-Chat [27]	73.46	57.11	73.94	72.26
RetSTA-7B-Zero	91.93	87.89	92.74	92.77
RetSTA-7B	**93.35**	**89.83**	**94.19**	**94.30**
>10B Large Language Models				
Qwen2.5-72B-Instruct [28]	81.97	72.39	84.34	83.21
GLM-4-Plus [11]	83.00	74.21	85.68	84.41
DeepSeek-V3 [19]	82.92	74.75	83.97	84.35

of the inherent logic and structure of text reports. It is noteworthy that despite the implementation of English prompts with explicit language constraints for English responses, Qwen2.5-7B-Instruct [28] generated outputs containing substantial Chinese content consistently. This linguistic inconsistency significantly compromised the model's performance in standardized evaluation scenarios.

Our method, RetSTA-7B, outperforms all compared models, including LLMs with significantly larger parameter sizes than 7B. It demonstrates that the proposed paradigm achieves a high-performance standardization model at a relatively low cost, while further validating that RetSTA-7B not only fully understands the correspondence between original reports and standard terminology, but also accurately masters complex expressions encountered in real clinical scenarios.

For Chinese reports, the results are shown in Table 2. RetSTA-7B significantly outperforms comparison models of similar size as well as those with substantially larger parameter sizes than 7B, demonstrating its strong domain adaptation capability. Notably, RetSTA-7B-Zero also demonstrates impressive performance, indicating that even simple data augmentation, which simulates diverse scenarios in clinical diagnostic reports, can achieve high performance in the task of standardization. This suggests that the proposed data augmentation strategy captures the variability and complexity of real clinical scenarios effectively, providing a cost-efficient approach to model training. Furthermore, the performance gap between RetSTA-7B and RetSTA-7B-Zero highlights the importance of incorporating large-scale bilingual standardized data for further enhancing model capabilities.

Table 3. The impact of data augmentation on RetSTA-7B-Zero's standardization performance.

	Samples	**BLEU-1**	**BLEU-4**	**METEOR**	**ROUGE-L**
w/ Augmentation	2000	87.11	78.84	87.56	87.82
w/o Augmentation	2000	87.49	79.51	88.08	88.16
w/ Augmentation	10000	91.93	87.89	92.74	92.77

Table 4. Comparison of test results from fine-tuning using Chinese reports only versus bilingual reports.

	BLEU-1	**BLEU-4**	**METEOR**	**ROUGE-L**
RetSTA-7B-Chinese	93.43	89.84	94.24	94.29
RetSTA-7B	93.35	89.83	94.19	94.30

3.3 Ablation Study

In Table 3, we investigate the impact of data augmentation strategies on the standardization performance of the fine-tuned Qwen2.5-7B-Instruct model. We present results under three experimental conditions: fine-tuning using 400 report pairs expanded to 2000 through data augmentation, fine-tuning directly using 2000 report pairs, and the results of RetSTA-7B-Zero. The experiments further validate that simulating complex scenarios in real clinical diagnostic environments to achieve the same data volume can achieve performance comparable to fine-tuning with real data, while significantly reducing costs. This finding provides strong support for efficient and cost-effective model fine-tuning. In Table 4 and Table 5, we observe that incorporating English translations into the fine-tuning data does not compromise the model's standardization performance for Chinese diagnostic reports, and vice versa. This finding indicates that the model exhibits strong independence and compatibility when handling bilingual data.

4 Conclusion

In this study, we construct a bilingual fundus standard terminology that integrates clinical standard terms and commonly used descriptions, and propose two LLMs for standardization: RetSTA-7B-Zero and RetSTA-7B. RetSTA-7B-Zero effectively handles complex expressions in clinical diagnostic processes,

Table 5. Comparison of test results from fine-tuning using English reports only versus bilingual reports.

	BLEU-1	**BLEU-4**	**METEOR**	**ROUGE-L**
RetSTA-7B-English	92.55	89.56	86.14	90.12
RetSTA-7B	92.69	89.66	86.04	90.08

demonstrating exceptional performance in standardization tasks. RetSTA-7B is more powerful, exhibiting superior domain adaptability, significantly outperforming other LLMs in standardization task. For future work, we plan to leverage RetSTA-7B to support downstream tasks, including training text-image foundation models and multimodal LLMs, as well as extending its application to domains beyond ophthalmology.

Acknowledgments. This research work is supported by Beijing Natural Science Foundation (Grant No.IS23112) and Beijing Institute of Technology Research Fund Program for Young Scholars (Grant No.6120220236).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
3. Cai, C.X., Halfpenny, W., Boland, M.V., Lehmann, H.P., Hribar, M., Goetz, K.E., Baxter, S.L.: Advancing toward a common data model in ophthalmology: Gap analysis of general eye examination concepts to standard observational medical outcomes partnership (omop) concepts. Ophthalmology Science **3**(4), 100391 (2023)
4. Chen, X., Xu, P., Li, Y., Zhang, W., Song, F., He, M., Shi, D.: Chatffa: An ophthalmic chat system for unified vision-language understanding and question answering for fundus fluorescein angiography. Iscience **27**(7) (2024)
5. Chen, X., Zhang, W., Xu, P., Zhao, Z., Zheng, Y., Shi, D., He, M.: Ffa-gpt: an automated pipeline for fundus fluorescein angiography interpretation and question-answer. npj Digital Medicine **7**(1), 111 (2024)
6. Chen, X., Zhang, W., Zhao, Z., Xu, P., Zheng, Y., Shi, D., He, M.: Icga-gpt: report generation and question answering for indocyanine green angiography images. British Journal of Ophthalmology (2024)
7. Chen, X., Zhao, Z., Zhang, W., Xu, P., Gao, L., Xu, M., Wu, Y., Li, Y., Shi, D., He, M.: Eyegpt: Ophthalmic assistant with large language models. arXiv preprint arXiv:2403.00840 (2024)
8. Domalpally, A., Fickweiler, W., Levine, S.R., Goetz, K.E., VanderBeek, B.L., Lee, A., Sundstrom, J.M., Markel, D., Sun, J.K.: Data harmonization, standardization, and collaboration for diabetic retinal disease (drd) research: report from the 2024 mary tyler moore vision initiative workshop on data. Translational Vision Science & Technology **13**(10), 4–4 (2024)
9. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
10. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. science **315**(5814), 972–976 (2007)

11. GLM, T., Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Zhang, D., Rojas, D., Feng, G., Zhao, H., et al.: Chatglm: A family of large language models from glm-130b to glm-4 all tools. arXiv preprint arXiv:2406.12793 (2024)
12. Halfpenny, W., Baxter, S.L.: Towards effective data sharing in ophthalmology: data standardization and data privacy. Current opinion in ophthalmology **33**(5), 418–424 (2022)
13. Hoffmann, K., Nesterow, I., Peng, Y., Henke, E., Barnett, D., Klengel, C., Gruhl, M., Bartos, M., Nüßler, F., Gebler, R., et al.: Streamlining intersectoral provision of real-world health data: a service platform for improved clinical research and patient care. Frontiers in Medicine **11**, 1377209 (2024)
14. Hoffmann, K., Peng, Y., Schlosser, T., Stolze, G., Langner, H., Susky, M., Meyer, T., Ritter, M., Kowerko, D., Kakkassery, V., et al.: Towards standardizing ophthalmic data for seamless interoperability in eye care. In: German Medical Data Sciences 2024, pp. 139–145. IOS Press (2024)
15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
16. Jin, K., Ye, J.: Artificial intelligence and deep learning in ophthalmology: Current status and future perspectives. Advances in ophthalmology practice and research **2**(3), 100078 (2022)
17. Li, J., Guan, Z., Wang, J., Cheung, C.Y., Zheng, Y., Lim, L.L., Lim, C.C., Ruamviboonsuk, P., Raman, R., Corsino, L., et al.: Integrated image-based deep learning and language models for primary diabetes care. Nature medicine **30**(10), 2886–2896 (2024)
18. Li, Z., Song, D., Yang, Z., Wang, D., Li, F., Zhang, X., Kinahan, P.E., Qiao, Y.: Visionunite: A vision-language foundation model for ophthalmology enhanced with clinical knowledge. arXiv preprint arXiv:2408.02865 (2024)
19. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
20. Pal, M.S.A., Sankarasubbu, M.: Openbiollms: Advancing open-source large language models for healthcare and life sciences (2024)
21. Sheehan, J., Hirschfeld, S., Foster, E., Ghitza, U., Goetz, K., Karpinski, J., Lang, L., Moser, R.P., Odenkirchen, J., Reeves, D., et al.: Improving the value of clinical research through the use of common data elements. Clinical Trials **13**(6), 671–676 (2016)
22. Shweikh, Y., Sekimitsu, S., Boland, M.V., Zebardast, N.: The growing need for ophthalmic data standardization. Ophthalmology Science **3**(1) (2023)
23. Tan, T.F., Thirunavukarasu, A.J., Jin, L., Lim, J., Poh, S., Teo, Z.L., Ang, M., Chan, R.P., Ong, J., Turner, A., et al.: Artificial intelligence and digital health in global eye health: opportunities and challenges. The Lancet Global Health **11**(9), e1432–e1443 (2023)
24. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
25. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data **3**(1), 1–9 (2016)

26. Xie, Y., Wu, J., Tu, H., Yang, S., Zhao, B., Zong, Y., Jin, Q., Xie, C., Zhou, Y.: A preliminary study of o1 in medicine: Are we closer to an ai doctor? arXiv preprint arXiv:2409.15277 (2024)
27. Yang, A., Xiao, B., Wang, B., Zhang, B., Bian, C., Yin, C., Lv, C., Pan, D., Wang, D., Yan, D., et al.: Baichuan 2: Open large-scale language models. arXiv preprint arXiv:2309.10305 (2023)
28. Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al.: Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115 (2024)
29. Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Wang, G., Li, H., Zhu, J., Chen, J., et al.: Yi: Open foundation models by 01. ai. arXiv preprint arXiv:2403.04652 (2024)
30. Zhang, K., Zeng, S., Hua, E., Ding, N., Chen, Z.R., Ma, Z., Li, H., Cui, G., Qi, B., Zhu, X., et al.: Ultramedical: Building specialized generalists in biomedicine. arXiv preprint arXiv:2406.03949 (2024)
31. Zhong, T., Liu, Z., Pan, Y., Zhang, Y., Zhou, Y., Liang, S., Wu, Z., Lyu, Y., Shu, P., Yu, X., et al.: Evaluation of openai o1: Opportunities and challenges of agi. arXiv preprint arXiv:2409.18486 (2024)