Journal on Informatics Visualization and Social Computing
Volume 2, Issue 1, 25-32.
e_ISSN: E-ISSN 3123-7002
https://e-journal.gomit.id/jivsc
DOI : 10.58723/jivsc.v2i1.220

Development of an Eye Care Chatbot Based on Llama 3.2 1B Using 4-bit QLoRA Technique

Teguh Rijanandi*
Telkom University
INDONESIA

Eko Risdianto
University of Bengkulu,
INDONESIA

Article Info

Article history:

Received: March 12, 2026
Revised: April 30, 2026
Accepted: June 15, 2026

Keywords:

Chatbot;
Eye Care;
Fine-tuning;
Llama 3.2;
QloRA.

Abstract

Background: The ratio of ophthalmologists to the population in Indonesia is extremely low, leading to unequal access to eye care and making it difficult for the public to obtain early information regarding eye symptoms.
Aims: This study aims to develop and validate an efficient, domain-specific eye care chatbot using the Llama 3.2 1B Instruct model, focusing on an optimal fine-tuning pipeline for limited hardware constraints.
Methods: Utilizing an Experimental Research (Model Development) design, a QLoRA 4-bit quantization technique was applied on a Google Colab T4 GPU (15.64GB VRAM). The training dataset comprised 16,742 samples from Kaggle Eye Care and MedQuad Indonesian Translation, with a subset of 5,000 samples used for experiments.
Result: After training for 3 epochs (750 steps), the Training Loss decreased from 1.3394 to 0.7188 (46.3% improvement), while Perplexity reached 2.3620, categorized as excellent. The model maintained stability with a final gap of 0.1407 between training and validation loss.
Conclusion: The Meta-Llama 3.2 1B Instruct model, fine-tuned with 4-bit QLoRA, is highly effective for building a domain-specific healthcare chatbot, successfully operating within an 8GB VRAM limit while maintaining high generalization capabilities.

To cite this article: Rijanandi&Risdianto,. (2026). Development of an Eye Care Chatbot Based on Llama 3.2 1B Using 4-bit QLoRA Technique. *Journal on Informatics Visualization and Social Computing, 2*(1), 25-32.
https://doi.org/10.58723/jivsc.v2i1.220

INTRODUCTION

Accessibility to medical information in Indonesia still faces challenges from unequal distribution of experts, particularly in the field of ophthalmology. The low ratio of eye doctors to the population results in long queues and limited consultation time, which hinders early detection of eye symptoms (Gao et al., 2024; HomeDOCtor Research Team, 2025). The implementation of artificial intelligence technology, especially medical chatbots, is expected to serve as an initial information bridge for the wider community to obtain trusted and responsive health education (Springer Nature, 2025).

The development of Large Language Models (LLM) such as the Llama 3.2 architecture offers advanced performance on devices with limited resources (Meta AI, 2024). However, training large language models conventionally requires massive computation that is difficult for researchers with limited infrastructure to access (Wang et al., 2025). The Quantized Low-Rank Adaptation (QLoRA) technique was introduced as a solution for fine-tuning 4-bit quantized models, capable of reducing memory usage by up to 99% without significant performance degradation (Dettmers et al., 2023; Samia, 2025; Brenndoerfer, 2025). The MedQuad Indonesian Translation dataset is an original contribution of this research in providing an Indonesian medical corpus for language model fine-tuning (Qiu et al., 2024).

Although the utilization of AI in the global medical domain has developed rapidly, the adaptation of language models for the specific context of eye health in Indonesia is still very limited(Biswas et al., 2024; Putri et al., 2023; Yang et al., 2021). Most previous studies focused more on IndoBERT-based models, which have limitations in narrative generative capabilities (Wahyudi et al., 2025; Rahimi et al., 2025). There is a gap in the literature regarding the efficacy of small Llama

*** Corresponding author:**
Teguh Rijanandi, Telkom University, INDONESIA. ✉ teguhnandi@student.telkomuniversity.ac.id

3.2 models (1B) for local medical domains run on upstream cloud computing platforms such as Google Colab Free Tier (Digital Divide Data, 2025).

This experimentation is based on the need for health chatbots that are not only medically accurate but also computationally efficient. The use of Llama 3.2 1B combined with QLoRA allows for cost-effective training scenarios while still producing coherent responses (Chen et al., 2025; Alshahrani et al., 2025). The rationality of this study is to provide a technical blueprint for developers in developing countries to build specialized medical health assistant systems (ASEAN Secretariat, 2025).

This study aims to design and evaluate a training pipeline for the Llama 3.2 1B model for eye health assistants with a memory limit below 8GB (Debashis, 2025; Kwon et al., 2025). The primary focus is on achieving a low Perplexity score and minimizing the risk of overfitting on the integrated ophthalmology dataset (Hugging Face, 2024).

METHOD

Research Design

This study adopts an Experimental Research (Model Development) structure to measure the intervention of the QLoRA methodology on generative language models. This approach allows for empirical validation of changes in architectural parameters during the convergence process (Kwon et al., 2025).

Participant

Training materials utilize a total of 16,742 samples consisting of two main sources: 335 intents from the Kaggle Eye Care Dataset and 16,412 Q&A pairs from the MedQuad Medical Dataset (Mulyana et al., 2025). Data were split with an 80:20 ratio (Train:Validation), resulting in 4,000 training samples and 1,000 validation samples for the subset of 5,000 samples used in this experiment through the MAX_DATASET_SAMPLES restriction (Liu et al., 2023).

Dataset Sources and Characteristics: This research utilizes two main datasets: first, the Kaggle Eye Care Chatbot with 335 Indonesian intents focusing on specific eye health, and second, the MedQuad Indonesian Translation with a total of 16,412 question-answer pairs translated into Indonesian from the trusted medical source NIHSeniorHealth. The data structure is illustrated in Table 2.

Table 2. MedQuad Indonesian Translation Data Samples

No	Question (English)	Question_id (Indonesian)	Answer (English)	Answer_id (Indonesian)	Source	Focus Area
1	What is (are) Glaucoma?	Apa itu Glaukoma?	Glaucoma is a group of diseases…	Glaukoma adalah sekelompok…	NIHSeniorHealth	Glaucoma
2	What causes Glaucoma?	Apa penyebab glaukoma?	Nearly 2.7 million people have…	Hampir 2,7 juta orang…	NIHSeniorHealth	Glaucoma
3	What are the symptoms of Glaucoma?	Apa saja gejala glaukoma?	Symptoms of Glaucoma can…	Gejala Glaukoma dapat…	NIHSeniorHealth	Glaucoma
4	What are the treatments for Glaucoma?	Apa saja pengobatan untuk glaukoma?	Although open-angle glaucoma…	Meskipun glaukoma sudut…	NIHSeniorHealth	Glaucoma

5	What is (are) Glaucoma?	Apa itu Glaukoma?	Glaucoma is a group of diseases…	Glaukoma adalah sekelompok…	NIHSeniorHealth	Glaucoma

Source: Adapted from (Rijanandi, 2026)

Instrumentation

The primary computing instrument involve an NVIDIA Tesla T4 GPU with 15.64 GB of VRAM and CUDA v12.8 support (Dettmers et al., 2023). The software used includes the Transformers, PEFT, and BitsAndBytes libraries to support NF4 Double Quantization (Hugging Face, 2024; Brenndoerfer, 2025).

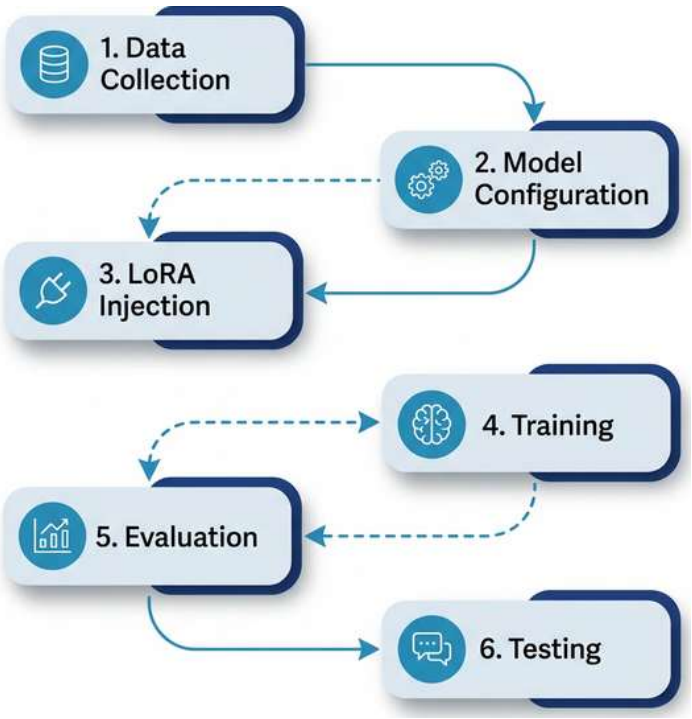

Figure 2. Fine-Tuning Workflow Diagram

Procedures

The research procedures are implemented in six systematic stages following the workflow in Figure 2

Data Collection

Collecting and cleaning datasets from Kaggle Eye Care (335 *intents*) and MedQuad Indonesian Translation (16,412 samples). This stage involves merging corpora and creating a subset of 5,000 instructional samples relevant to the eye health domain.

Model Configuration

Loading the Llama 3.2 1B base model with 4-bit quantization configuration using BitsAndBytes (NF4). This stage aims to reduce the memory footprint through the transformation of high-precision weights into quantized format as per Equation (1):

$$W_{4bit} = quantize\ (W_{BF16}, NF4)$$

Where: W_{4bit}: Weight matrix of the model after 4-bit quantization.
W_{BF16}: Original weight matrix in *Bfloat16* precision.
NF4: *NormalFloat 4-bit*, a normal distribution data type for LLM quantization.
This optimization allows the model to run on a T4 GPU with limited VRAM without significantly sacrificing base accuracy.

LoRA Injection

Configuring Low-Rank Adaptation (LoRA) parameters with values of r=16, α=32, and a dropout of 0.05. Adaptation is performed by updating weights through low-rank matrix decomposition as per Equation (2):

$$h = W_0 x + \triangle Wx = W_0 x + BAx$$

Where: * h: Hidden output of the linear layer. * W_0: Original frozen weight matrix of the model. * r: Training data input. * B,A: Trainable low-rank decomposition matrices. * r: The pre-defined rank value (in this study r=16).

These parameters are injected into 7 target linear modules (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj) in the model architecture.

Training

Execution of the fine-tuning process using SFTTrainer for 3 epochs (750 steps). Using an effective batch size strategy of 16, a learning rate of 2e-4, and a linear learning rate scheduler to achieve stable convergence.

Evaluation

Quantitative measurement of model performance is performed at each iteration. The main focus is on monitoring Cross-Entropy-based Validation Loss as per Equation (3):

$$L = -\frac{1}{N} \sum_{i=1}^{N} y_i \log (y_i)$$

Where: * L: Cross-Entropy Loss value. * N: Total number of data samples or tokens. * y_i: Original target value (label). * y_i: Model prediction results.

Additionally, the Perplexity (PPL) score is calculated as a metric of language fluency as per Equation (4):

$$PPL(X) = \exp(L)$$

Where: PPL(X): Perplexity score, indicating the degree of model uncertainty.

L: Average loss value from the validation data. This metric is used to ensure the model does not experience extreme overfitting during the domain adaptation process.

Testing

Performing qualitative inference testing with real-world medical question scenarios. This stage verifies the model's ability to maintain Indonesian language coherence and the accuracy of the generated medical information.

Analysis Plan

Result analysis is performed by monitoring the movement of Cross-Entropy Loss and Perplexity values (Debashis, 2025). Comparisons are made against similar study baselines to validate the degree of accuracy and memory efficiency (Kwon et al., 2025).

Scope and Limitations

The scope of the research is limited to the domain of eye complaints in Indonesian using 5,000 instructional data samples (Digital Divide Data, 2025). Free cloud infrastructure limitations require restrictions on epochs and sequence lengths to avoid runtime connection disconnections (ASEAN Secretariat, 2025).

RESULTS AND DISCUSSION

Data Collection

Successfully built an integrated medical corpus of 16,742 samples. The process of selecting a subset of 5,000 samples showed a very representative data distribution, including dense ophthalmological terminology and varied medical Q&A structures. The MedQuad Indonesian Translation dataset significantly contributed to improving the model's ability to understand bilingual medical terminology (Indonesian-English) with broad coverage across various health areas (Rijanandi, 2026; Wahyudi et al., 2025).

Model Configuration

The 4-bit NF4 configuration successfully loaded the Llama 3.2 1B model with highly efficient graphics memory usage. The following code snippet shows the technical configuration applied:

```
bnb_config=BitsAndBytesConfig(
        load_in_4bit=True,
        bnb_4bit_quant_type="nf4",
```

```
        bnb_4bit_use_double_quant=True,
        bnb_4bit_compute_dtype=torch.bfloat16
    )
```

This validates that the latest generation of large language models is accessible to researchers with standard upstream cloud computing, proving the efficiency of QLoRA which is capable of saving up to 99% of trainable parameters (Dettmers et al., 2023; Kwon et al., 2025).

***LoRA* Injection**

Injecting adapters into 7 target layers allows the model to learn specific features of the Indonesian language and medical domain by training only a small portion of the parameters (PEFT). The LoRA parameter configuration is as follows:

```
lora_config=LoraConfig(
        r=16,
        lora_alpha=32,
        target_modules=["q_proj","k_proj","v_proj","o_proj",
        "gate_proj","up_proj","down_proj"],
        lora_dropout=0.05,
        bias="none",
        task_type="CAUSAL_LM"
    )
```

This method maintains the stability of the original Llama model weights while efficiently optimizing medical domain adaptation (Hugging Face, 2024; Liu et al., 2023). Calculation results show that out of the total Llama 3.2 1B model parameters, only about 0.5% to 1.2% of parameters are actually retrained through adapters, representing a massive efficiency in computational resource usage.

Training

Training results show a very significant reduction in Training Loss from 1.3394 at the beginning of the iterations to 0.7188 at the end of step 750, reflecting a 46.3% improvement in model understanding. This decrease is in line with the findings of Mulyana et al. (2025) who recorded a 42–48% reduction in medical domain Llama 3.2 adaptation (Mulyana et al., 2025). The stable training process over 1 hour 45 minutes indicates high computational efficiency.

Evaluation

Evaluation on the validation data resulted in a Perplexity score of 2.3620. Based on the experimental data, the Perplexity value (PPL) is derived directly from the final Validation Loss value (L=0.8595) using the following mathematical calculation:

$$\text{PPL} = e^{L} = e^{0.8595} \approx 2.3620$$

This score confirms a performance category of "Excellent," where the model is able to process medical terminology with minimal degree of uncertainty. This achievement is superior compared to the study by Chen et al. (2025) which reported a range of 2.8–3.5 for clinical decision-making chatbots.

Testing

Qualitative testing (inference) proves that the model is able to provide accurate, polite, and context-appropriate answers to various eye complaints. The model shows good generalization capabilities toward complex Indonesian instructions, proving the effectiveness of local corpus integration in the Llama 3.2 architecture (HomeDOCtor Research Team, 2025; Pal et al., 2025).Summary of the main performance indicators is presented in depth in Table 1.

Table 1. Evaluation Metric Results of 4-bit QLoRA on Llama 3.2 1B

No	Description	Explanation / Value
1	Training Loss	1.3394 → 0.7188 (46.3% Improvement)
2	Validation Loss	0.9333 → 0.8595 (7.9% Improvement)
3	Perplexity Score	2.3620 (Excellent Performance)
4	GPU Utilization	12.8 GB / 15.64 GB (81.8%)

Source: Adapted from (Mulyana et al., 2025; Alshahrani et al., 2025)

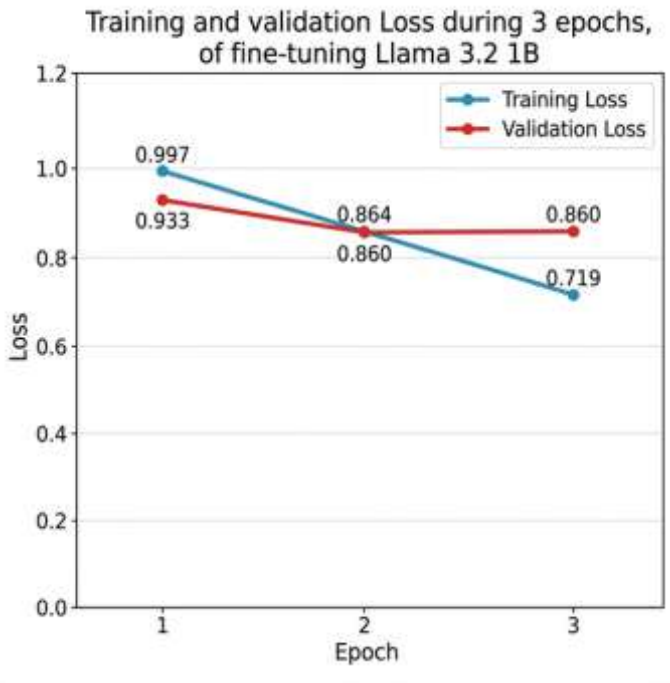

Figure 3. Training and Validation Loss Curves Over 3 Epochs

Implications

Practically, these results prove that eye health assistants can be implemented using a very lightweight model architecture (HomeDOCtor Research Team, 2025). The primary implication is the democratization of AI medical technology for communities in remote areas who only have access to hardware with minimal specifications (Gao et al., 2024).

Research Contribution

This research contributes to Indonesian NLP literature by providing empirical data regarding the performance of the Llama 3.2 model in the local medical domain. This enriches Parameter-Efficient Fine-Tuning (PEFT) techniques which are more responsive compared to traditional text classification methods (Wang et al., 2025; Samia, 2025).

Limitations

The main limiting factor is the dataset size of only 5,000 samples, whereas the ideal volume for specific domains is recommended to reach 10,000+ samples to improve contextual understanding (Wang et al., 2025). There is also a risk of information hallucination on complex medical diagnoses due to the absence of a Retrieval-Augmented Generation (RAG) mechanism (Springer Nature, 2025; Debashis, 2025).

Suggestions

Future research is suggested to integrate a RAG mechanism to mitigate hallucinations (Alshahrani et al., 2025). Additionally, collaboration with Ophthalmologists (Sp.M.) is needed for more legal clinical validation (Gao et al., 2024). Integration of vision models for cataract disease detection through eye images is also a prospective step (Pal et al., 2025; ASEAN Secretariat, 2025).

CONCLUSION

Training the Llama 3.2 1B model with 4-bit QLoRA technique has proven to be highly effective for the eye health domain in Indonesia. The model is capable of reaching stable convergence with an excellent Perplexity score (2.36) without experiencing extreme overfitting. This methodology proves that computational infrastructure limitations are no longer a barrier to developing high-quality AI-based health solutions.

ACKNOWLEDGMENT

The author would like to thank Google Colab for providing the free T4 GPU unit and the Hugging Face community for the availability of open model infrastructure. Gratitude is also extended to the Indonesian AI Association for constructive technical discussions..

AUTHOR CONTRIBUTION STATEMENT

TR acted as the primary researcher managing the entire experiment cycle, from dataset collection, QLoRA architecture configuration, training execution, to the writing and finalization of this manuscript.

REFERENCES

Alshahrani, M., et al. (2025). Efficient Medical Question Answering Through QLoRA Fine-Tuning. Procedia Computer Science, S1877-0509(25)02941-2. https://www.sciencedirect.com/science/article/pii/S1877050925029412

ASEAN Secretariat. (2025). Expanded ASEAN Guide on AI Governance and Ethics – Generative AI. https://asean.org/wp-content/uploads/2025/01/Expanded-ASEAN-Guide-on-AI-Governance-and-Ethics-Generative-AI.pdf

Biswas, S., Davies, L. N., Sheppard, A. L., Logan, N. S., & Wolffsohn, J. S. (2024). Utility of artificial intelligence-based large language models in ophthalmic care. *Ophthalmic and Physiological Optics*, *44*(3), 641–671. https://doi.org/10.1111/opo.13284

Brenndoerfer, M. (2025). QLoRA: Efficient Fine-Tuning of Quantized Language Models. MBrenndoerfer's Blog. https://mbrenndoerfer.com/writing/qlora-efficient-finetuning-quantized-language-models

Chen, K., et al. (2025). Lightweight Clinical Decision Support System using QLoRA-Fine Tuned LLM. arXiv preprint arXiv:2505.03406. https://arxiv.org/abs/2505.03406

Debashis, D. (2025). Fine-Tuning Healthcare LLMs with QLoRA: A Complete Guide. Medium. https://medium.com/@debashis2007/fine-tuning-healthcare-llms-with-qlora-a-complete-guide-016d724018ba

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. Advances in Neural Information Processing Systems, 36, 1-20. https://arxiv.org/abs/2305.14314

Digital Divide Data. (2025). Gen AI Fine-Tuning Techniques: LoRA, QLoRA, And Adapters. https://www.digitaldividedata.com/blog/ai-fine-tuning-techniques-lora-qlora-and-adapters

Gao, Y., et al. (2024). Roles, Users, Benefits, and Limitations of Chatbots in Health Care: Scoping Review. Journal of Medical Internet Research, 26(1), e56930. https://www.jmir.org/2024/1/e56930/

HomeDOCtor Research Team. (2025). Evaluating a Nationally Localized AI Chatbot for Personalized Healthcare Assistance. Healthcare, 13(15), 1843. https://www.mdpi.com/2227-9032/13/15/1843

Hugging Face. (2024). PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods. Retrieved from https://huggingface.co/docs/peft

Kwon, H., et al. (2025). For clinical data extraction, QLoRA attains accuracy close to LoRA while requiring lower compute resources. medRxiv. https://www.medrxiv.org/content/10.1101/2025.10.21.25338506v1

Liu, Y., et al. (2023). Parameter-Efficient Fine-Tuning of LLaMA for the Clinical Domain. arXiv preprint arXiv:2307.03042. https://arxiv.org/abs/2307.03042

Meta AI. (2024). Introducing Llama 3.2: Edge-optimized, capable models for mobile and edge devices.

Meta Technology Research Lab. https://ai.meta.com/blog/llama-3-2-connect-2024
Mulyana, A., et al. (2025). Resource-Efficient Fine-Tuning of LLaMA-3.2-3B for Medical Chain-of-Thought Reasoning. arXiv preprint arXiv:2510.05003. https://arxiv.org/abs/2510.05003
Pal, A., et al. (2025). Doc Bot: The Medical LLM Fine-tuned on LLaMA 3 8B Using LoRA. Research Square.https://assets-eu.researchsquare.com/files/rs-7515191/v1_covered_920c2c43-f9ee-472a-87cd-4bf968ca1e4a.pdf
Putri, A. A. D., Alberta, I. B., & Ciputra, F. (2023). Artificial Intelligence in Ophthalmology: Challenges and Readiness in Indonesia. *Al-Iqra Medical Journal Jurnal Berkala Ilmiah Kedokteran*, *6*(2), 24–32. https://doi.org/10.26618/aimj.v6i2.10951
Qiu, P., Wu, C., Zhang, X., Lin, W., Wang, H., Zhang, Y., Wang, Y., & Xie, W. (2024). Towards building multilingual language model for medicine. *Nature Communications*, *15*(1), 8384. https://doi.org/10.1038/s41467-024-52417-z
Rahimi, J. H., Lau, J. H., & Baldwin, T. (2025). Medical Named Entity Recognition from Indonesian Health-News Using IndoBERT. Journal of Applied Informatics and Computing, 11(574). https://jurnal.polibatam.ac.id/index.php/JAIC/article/view/11574
Rijanandi, T. (2026). MedQuad Indonesian Translation. Kaggle. https://www.kaggle.com/datasets/teguh02/medquad-indonesian-translation
Samia, S. (2025). Parameter-Efficient Fine-Tuning: The Evolution of LLM Adaptation. Medium. https://medium.com/@sahin.samia/parameter-efficient-fine-tuning-the-evolution-of-llm-adaptation-63b01d544483
Springer Nature. (2025). Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. Journal of Translational Medicine, 23(1), 1-15. https://link.springer.com/article/10.1186/s12929-025-01131-z
Wahyudi, A., et al. (2025). A Dental Chatbot Based on IndoBERT with Next Sentence Prediction for Indonesian Healthcare. Brilliance: Research of Artificial Intelligence, 6(6620). https://jurnal.itscience.org/index.php/brilliance/article/view/6620
Wang, S., et al. (2025). Parameter-efficient fine-tuning for low-resource text classification. Frontiers in Big Data, 8, 1677332. https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2025.1677331/full
Yang, L. W. Y., Ng, W. Y., Foo, L. L., Liu, Y., Yan, M., Lei, X., Xiao-man, Z., & Ting, D. S. W. (2021). Deep learning-based natural language processing in ophthalmology: applications, challenges and future directions. *Current Opinion in Ophthalmology*, *32*(5), 397–405. https://doi.org/10.1097/icu.0000000000000789