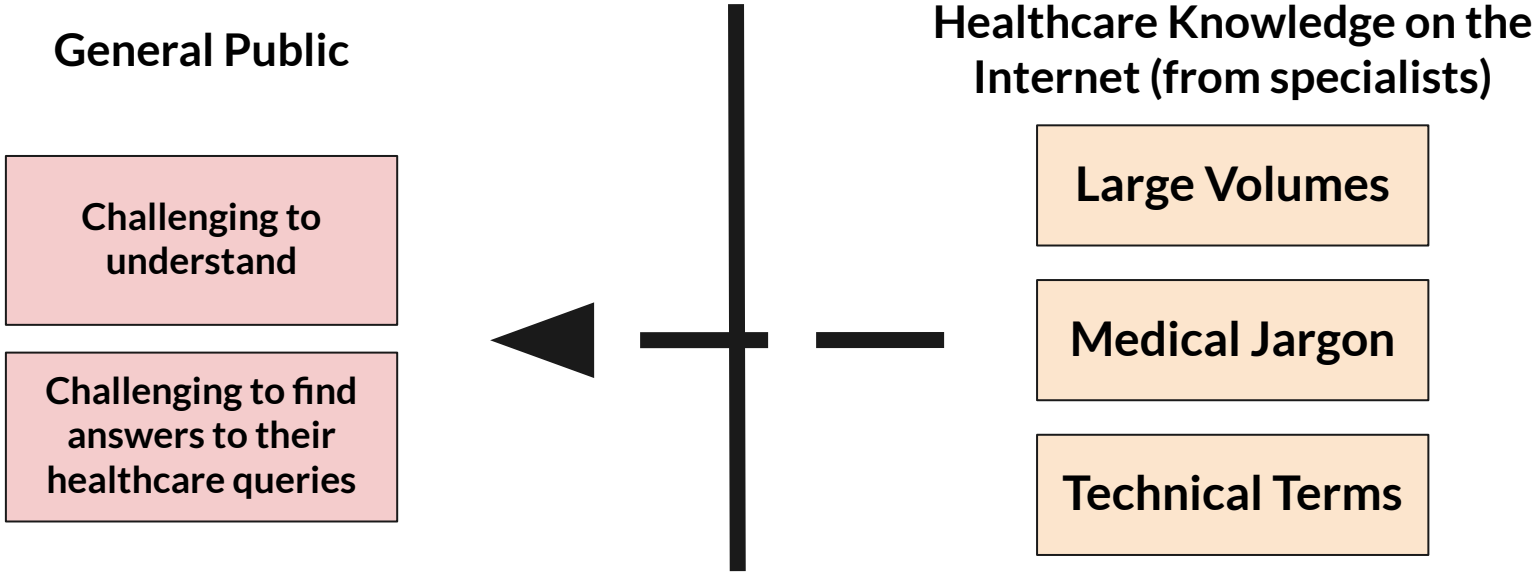
MediLiteQA: Tuning Smaller LMs for Healthcare QA

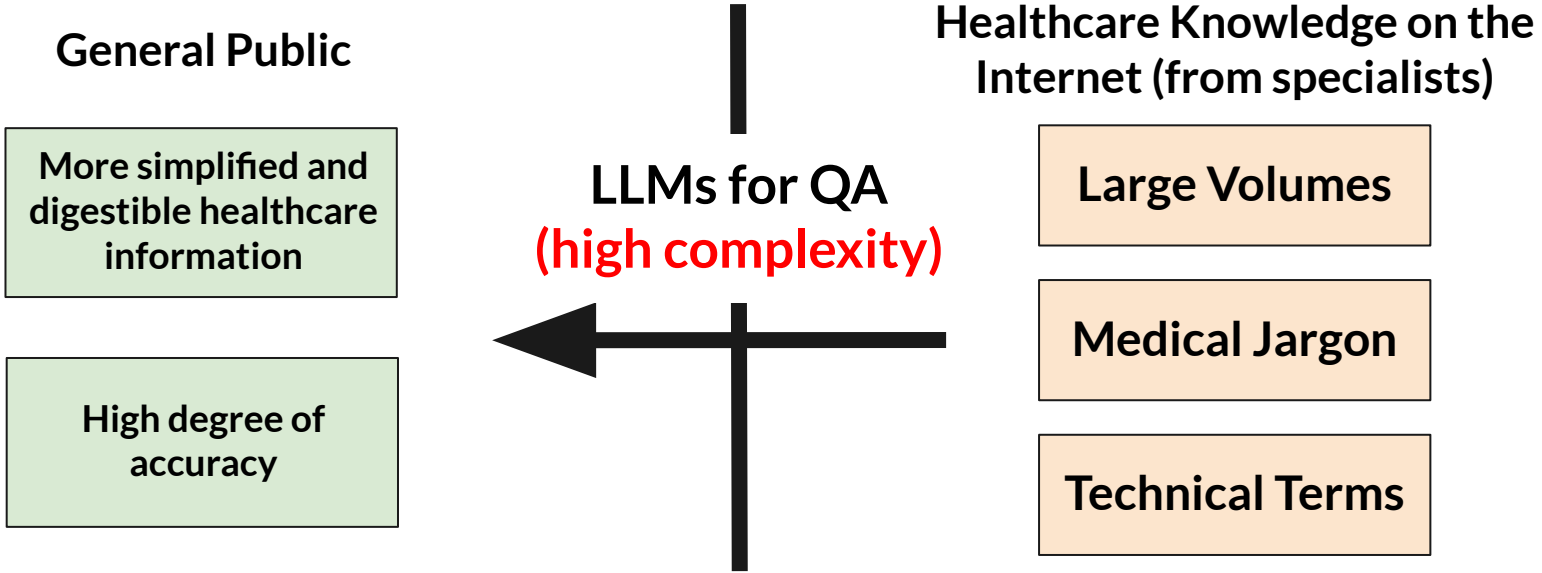
Group 11
Tay Kaiying, Roydon (A0271742E)
Ngu Jia Hao (A0272789H)
Chia Bing Xuan (A0259419R)

1. Introduction

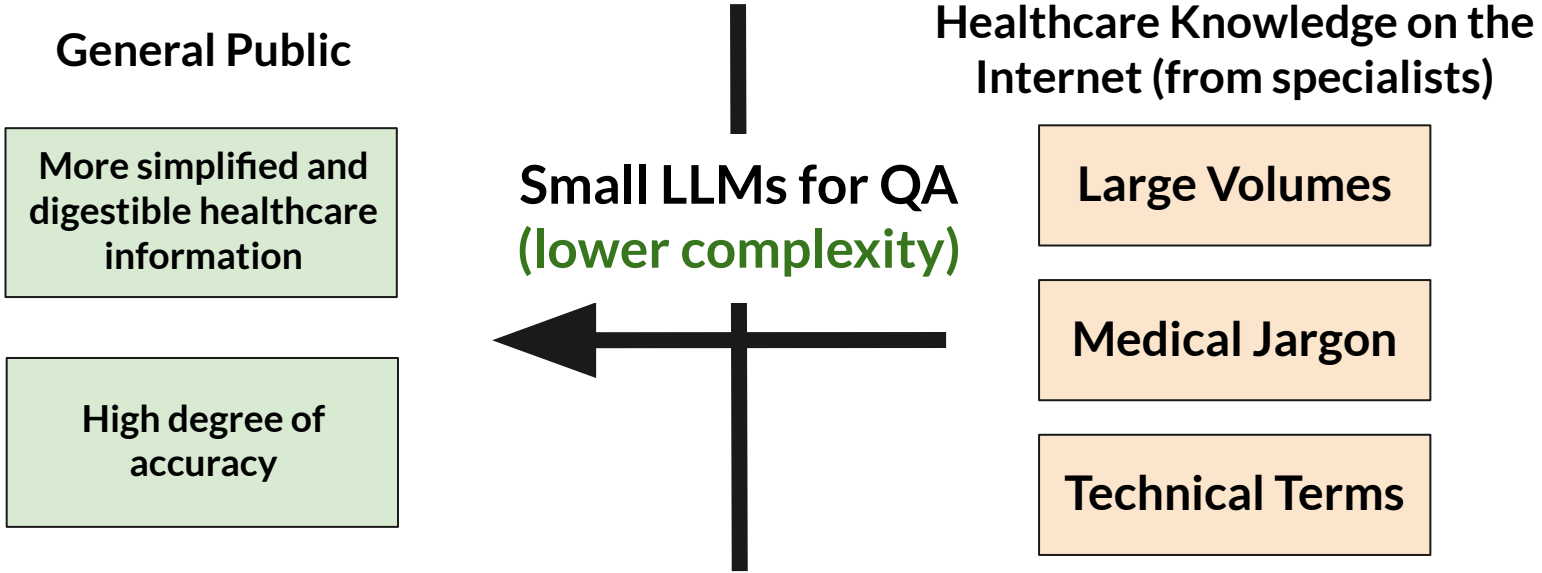
In healthcare, there is a knowledge gap between the layperson and the domain experts



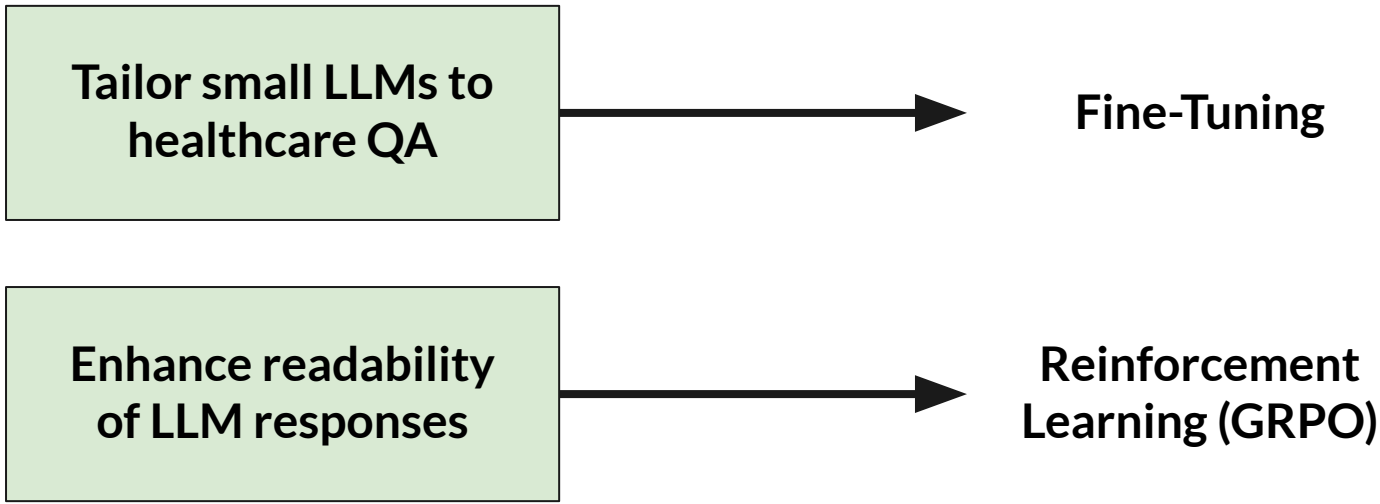
LLMs have bridged this mismatch, but they are expensive to train and store



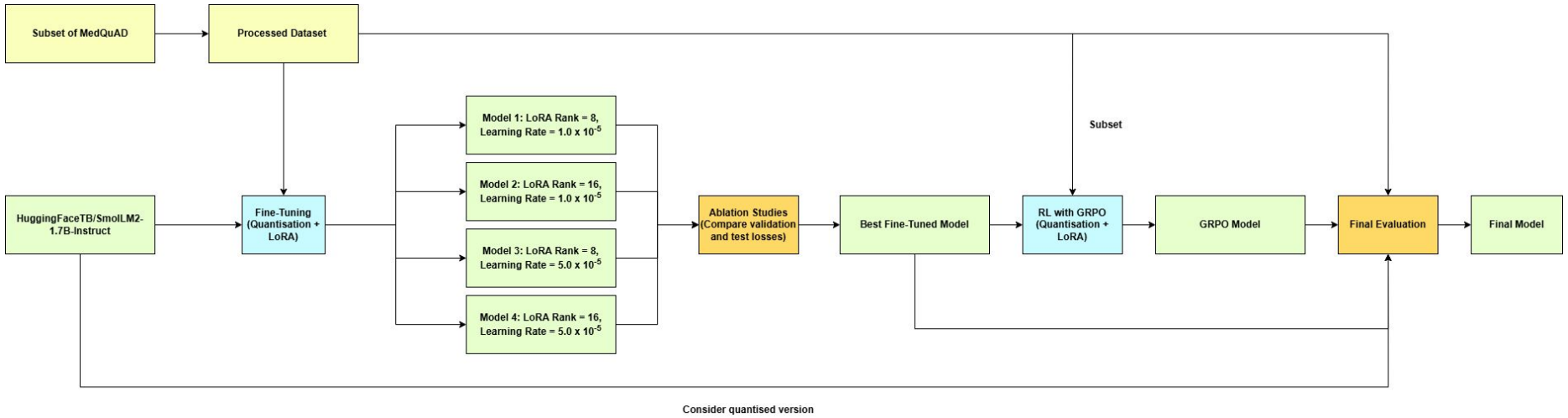
We aim to tune smaller LLMs for healthcare QA



Project Scope: Improving the performance of small LLMs in healthcare QA



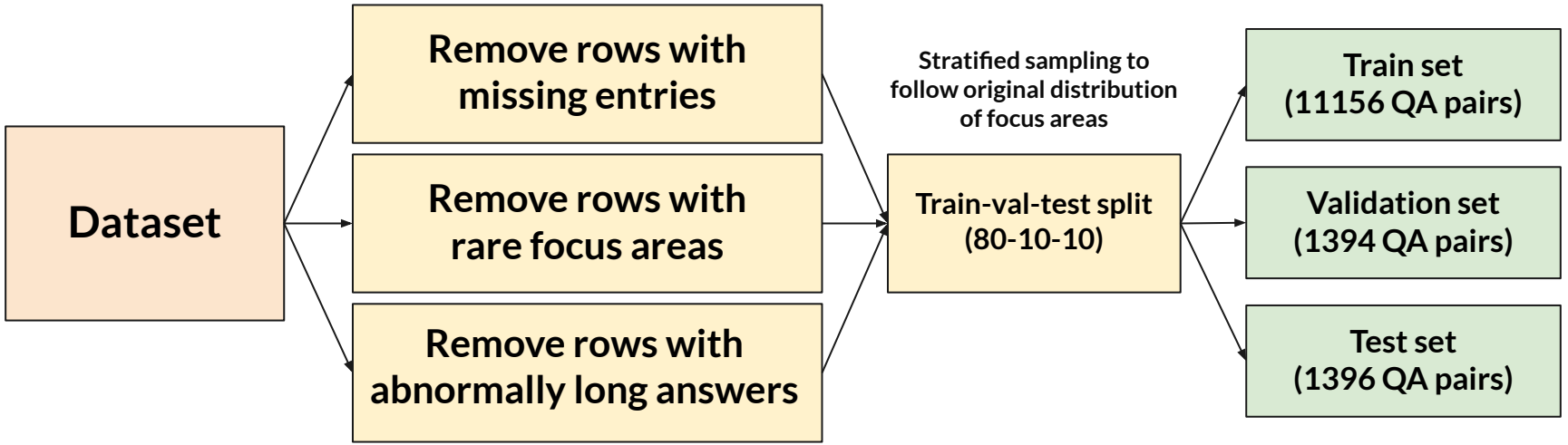
Workflow


Subset of MedQuAD
Processed Dataset
HuggingFaceTB/SmolLM2-1.7B-Instruct
Fine-Tuning (Quantisation + LoRA)
Model 1: LoRA Rank = 8, Learning Rate = 1.0 x 10-5
Model 2: LoRA Rank = 16, Learning Rate = 1.0 x 10-5
Model 3: LoRA Rank = 8, Learning Rate = 5.0 x 10-5
Model 4: LoRA Rank = 16, Learning Rate = 5.0 x 10-5
Ablation Studies (Compare validation and test losses)
Best Fine-Tuned Model
Subset
RL with GRPO (Quantisation + LoRA)
GRPO Model
Final Evaluation
Final Model
Consider quantised version

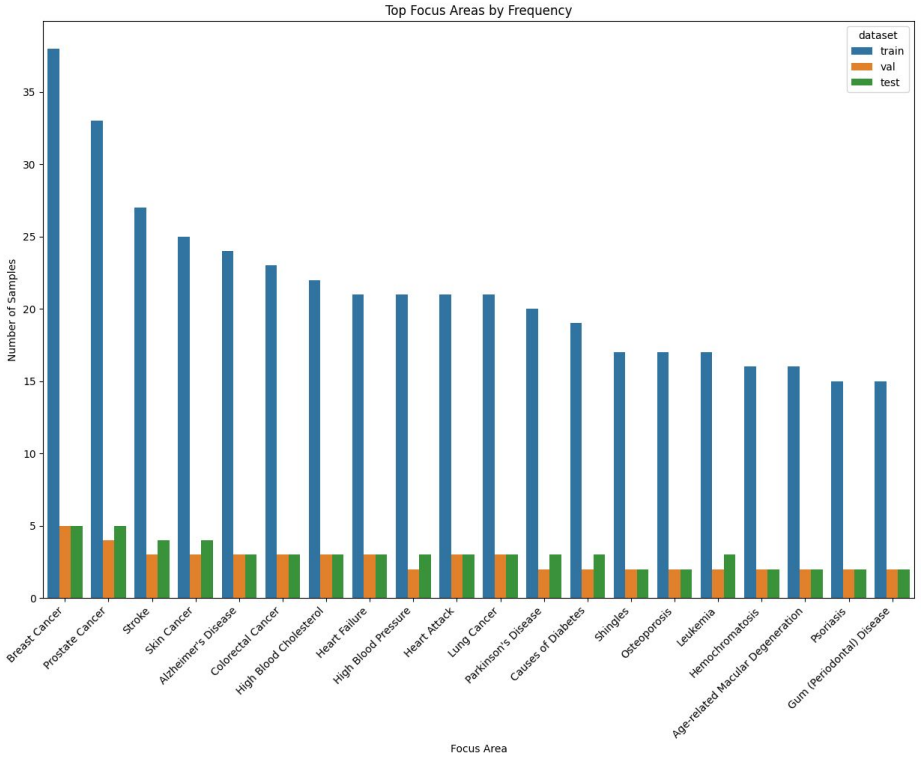
2. Data Cleaning and Processing

Dataset and Data Processing

Subset of Medical Question Answering Dataset (MedQuAD): 12 source websites, 14984 unique question-answer pairs, 5127 focus areas



Focus Area Composition of Train-Val-Test Split



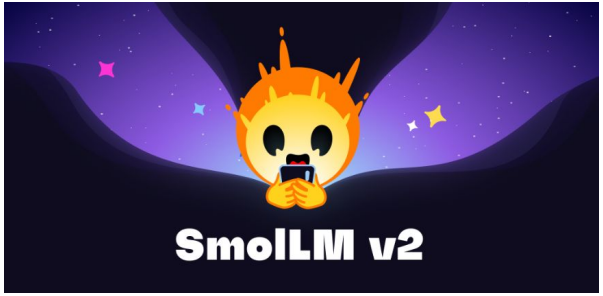
Dataset and Data Processing (GRPO)

Owing to computational constraints, we are unable to apply reinforcement learning on the entire training set. We shall use a subset of the training set for GRPO

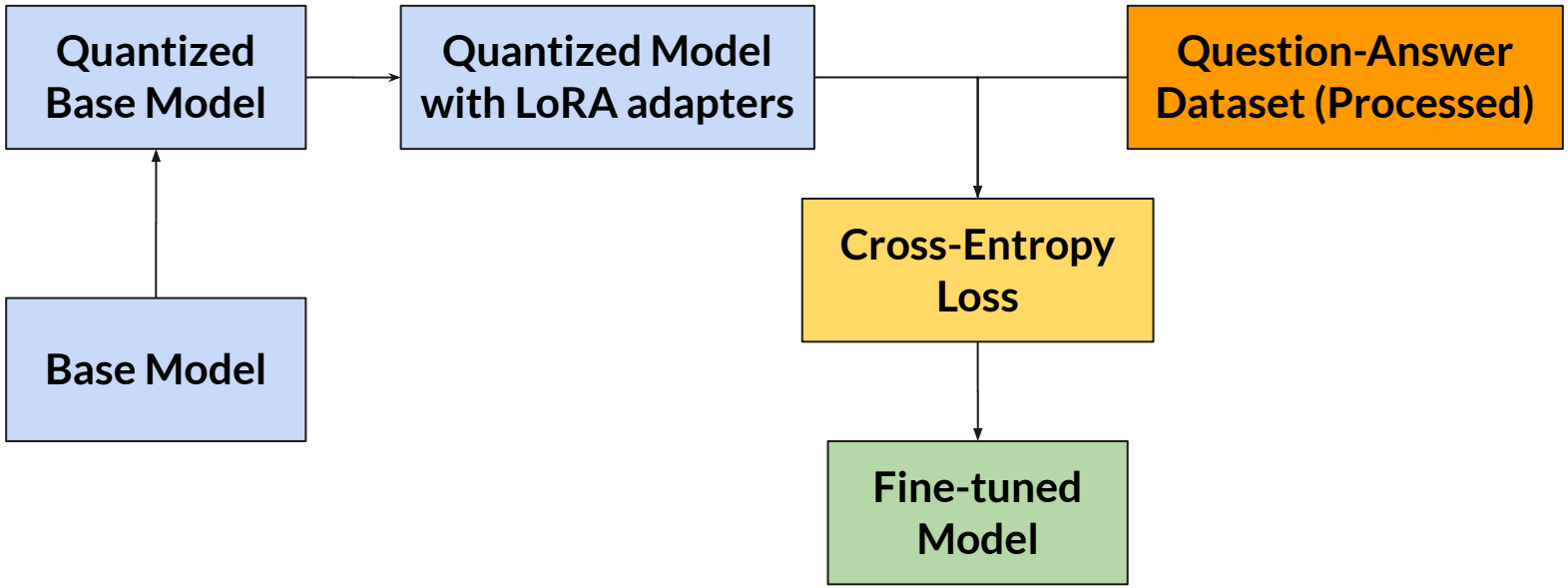
3. Supervised Fine-Tuning

Base Model chosen for fine-tuning

- Base Language model for fine-tuning: HuggingFaceTB/SmolLM2-1.7B-Instruct
- We train this small(er) language model for closed-book QA in the healthcare domain as they are computationally less expensive and can be run on a wider range of devices.
- We aim to improve the domain knowledge of this model in the healthcare domain through fine-tuning.



Fine-tuning model for closed-book QA



Ablation Studies

Model	LoRA Rank	Learning Rate	Validation Loss	Test Loss
1	8	1e-5	1.40	1.40
2	16	1e-5	1.42	1.42
3	8	5e-5	1.13	1.14
4	16	5e-5	1.14	1.14

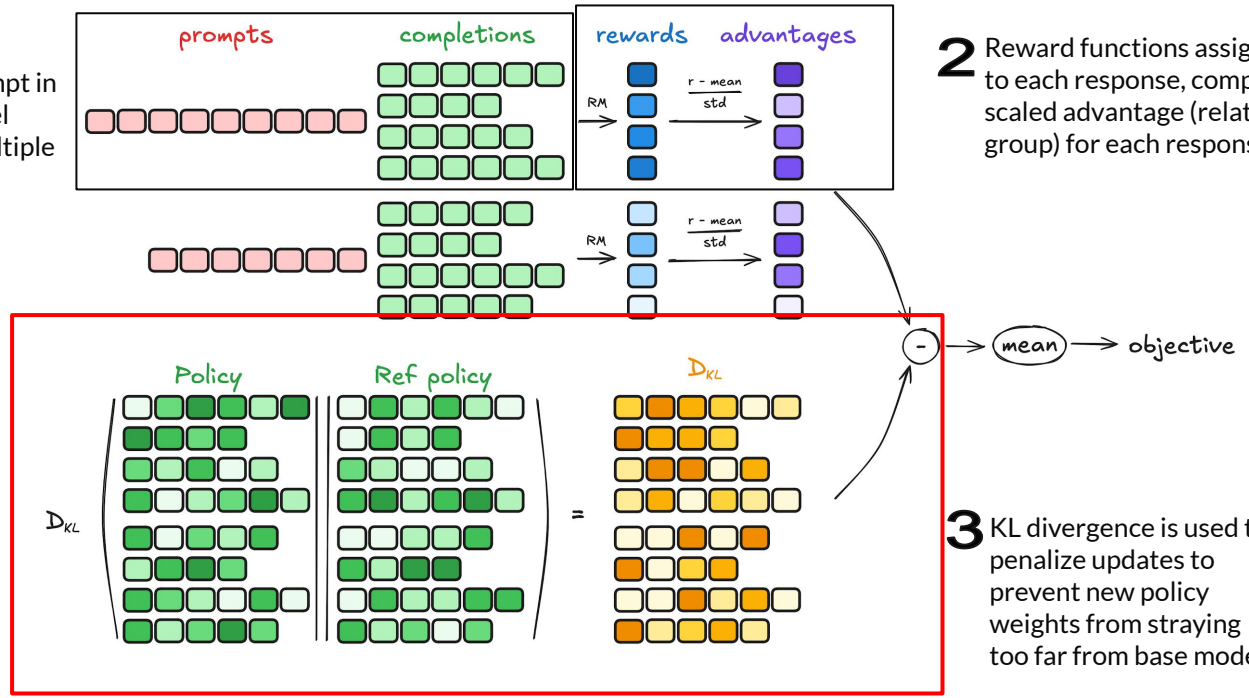
4. Reinforcement Learning (GRPO)

Reinforcement Learning with GRPO

- Why Reinforcement Learning?
  - Want the model to output responses that are easy to read and understand
- **Why GRPO?**
  - **Better than RL alternatives:** Eliminates the need for a value model (PPO)
    Also does not require a curated human preference dataset (DPO)

Reinforcement Learning with GRPO

1 For each prompt in dataset, model generates multiple responses (a group)



2 Reward functions assign scores to each response, compute the scaled advantage (relative to group) for each response

3 KL divergence is used to penalize updates to prevent new policy weights from straying too far from base model

GRPO Objective Function

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t}, \text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right] - \beta\mathbb{D}_{KL}\left[\pi_\theta||\pi_{ref}\right]\right\}$$

Reward functions chosen

Flesch Reading Ease Score: Readability scores using Average Sentence Length (ASL) and Average Syllables per Word (ASW)

Flesch Reading Ease = 206.835 – (1.015 x ASL) – (84.6 x ASW)

Soft Length-Based Punishment

Maximum reward for ideal length (30 <= word length <= 300)
Soft decay of reward towards -1 if the length of responses deviates on either side of ideal length bounds

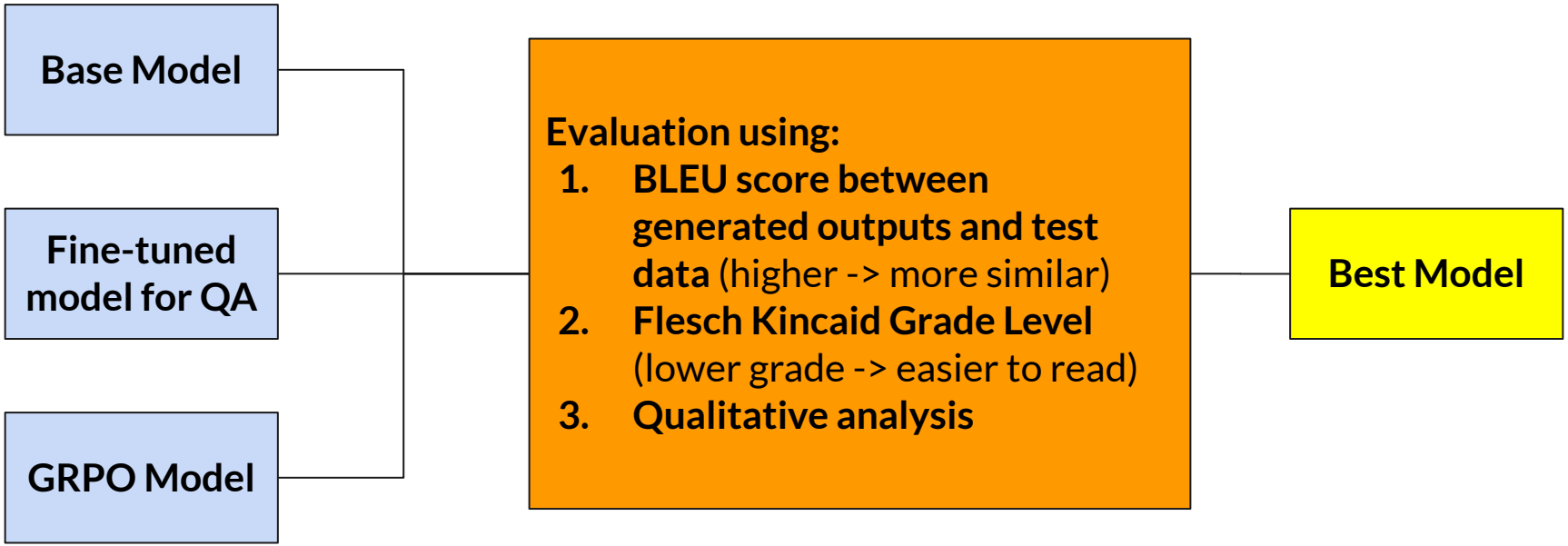
Reward functions chosen

Flesch Reading Ease Score: To get model to discover ways to make its responses more easy to read and understand (e.g. shorter sentences, simpler synonyms)

Soft Length-Based Punishment: To guide model to keep responses within an ideal length

5. Evaluation Results

Evaluation

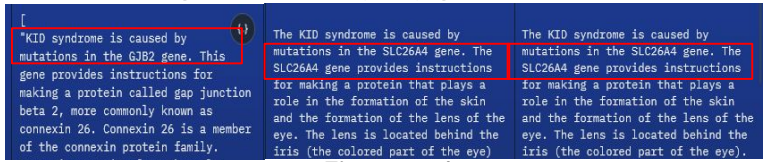

Base Model
Fine-tuned model for QA
GRPO Model
Evaluation using:
1. BLEU score between generated outputs and test data (higher -> more similar)
2. Flesch Kincaid Grade Level (lower grade -> easier to read)
3. Qualitative analysis
Best Model

Evaluation Results

	BLEU	**Flesch Kincaid Readability Grade**
Base Instruct Model	**0.04900**	12.3
Fine-tuned Model	0.02044	11.6
GRPO Trained Model	0.02777	**10.7**

Fine-tuned models still make mistakes

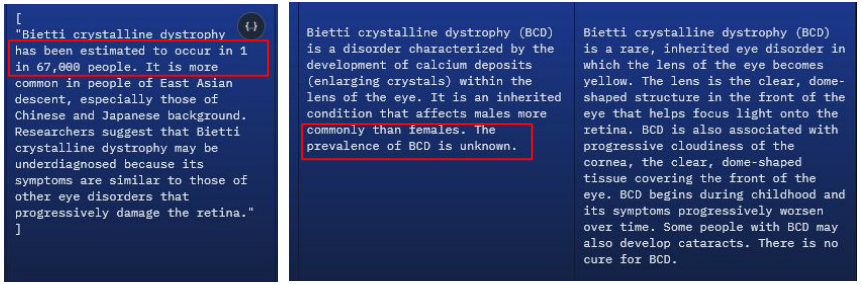
Question: What genetic mutation causes the keratitis-ichthyosis-deafness (KID) syndrome?



Answer | Fine-tuned | GRPO

Fine-tuned models got confused when conditions have similar symptoms (gave another gene that also causes deafness but not the one that causes KID syndrome)

Question: How many people are affected by Bietti crystalline dystrophy (BCD) ?

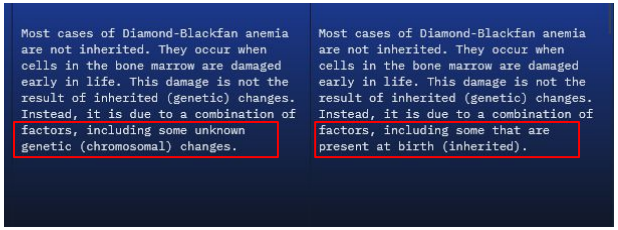


Answer | Fine-tuned | GRPO

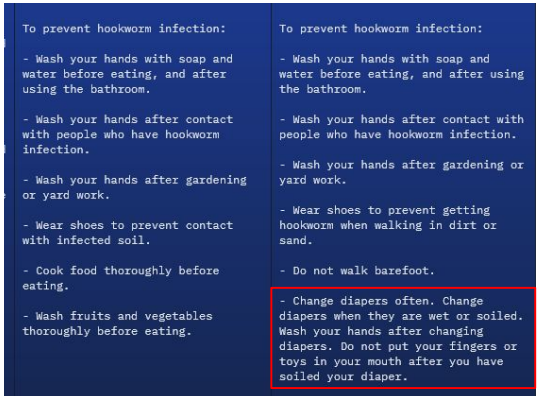
Fine-tuned models failed to learn some knowledge from the dataset (did not know the prevalence of BCD although present in data)

Dataset has insufficient training samples for imparting knowledge

Good and Bad examples from GRPO Training



Fine-tuned GRPO



In successful cases, GRPO trains model to improve word choices while keeping meaning the same
(genetic, chromosomal -> inherited)

However, GRPO caused model to hallucinate facts sometimes.
(hookworm infection should be unrelated to diapers)

GRPO works, but factuality may be affected

Possible Future Work

- **Synthetic Data Creation:** Large LLM can be used to generate quality synthetic data to provide more training samples for smaller LLM (knowledge distillation)
- **Unifying similar training samples:** Dataset contains varied answers for similar questions. These responses could be concatenated and summarized by teacher LLM -> reduces variety of answers for similar / same questions and help model learn better
- **Explore other reward models**: Using LLM-as-Judge for rewarding correctness, reward professional term exclusion

6. Q and A

References

- Abacha, A. B., & Demner-Fushman, D. (2019). A Question-Entailment Approach to Question Answering. *BMC Bioinformatics*, *20*(511). https://doi.org/10.1186/s12859-019-3119-4
- Allal, L. B., Lozhkov, A., Bakouch, E., Blázquez, G. M., Penedo, G., Tunstall, L., Marafioti, A., Kydlíček, H., Lajarín, A. P., Srivastav, V., Lochner, J., Fahlgren, C., Nguyen, X.-S., Fourrier, C., Burtenshaw, B., Larcher, H., Zhao, H., Zakka, C., Morlon, M., … Wolf, T. (2025). SmolLM2: When Smol Goes Big — Data-Centric Training of a Small Language Model. https://arxiv.org/html/2502.02737v1
- Cao, Y., Zhao, H., Cheng, Y., Shu, T., Chen, Y., Liu, G., Liang, G., Zhao, J., Yan, J., & Li, Y. (2024). Survey on Large Language Model-Enhanced Reinforcement Learning: Concept, Taxonomy, and Methods. *IEEE Transactions on Neural Networks and Learning Systems*, *36*(6), 9737 – 9757. https://arxiv.org/abs/2404.00282
- Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 10088 – 10115. https://doi.org/10.48550/arXiv.2305.14314
- Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., … Zhang, Z. (2025). Deepseek-R1: Incentivizing reasoning capability in LLMS via reinforcement learning. *arXiv.org*. https://arxiv.org/abs/2501.12948
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *The Tenth International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2106.09685
- Kincaid, J. P., Fishburne, R. P., Jr., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel*. Institute for Simulation and Training. https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1055&context=istlibrary

References

- Leo, C. (2024, July 16). *The Math Behind Multi-Head Attention in Transformers*. Towards Data Science. https://towardsdatascience.com/the-math-behind-multi-head-attention-in-transformers-c26cba15f625/
- Miller, J. (2024). Healthcare NLP: LLMs, Transformers, Datasets [Data set]. Kaggle. https://www.kaggle.com/datasets/jpmiller/layoutlm
- Qiu, W., Huang, T., Rullo, R., Kuang, Y., Maatouk, A., Ramos, S. R., & Ying, R. (2025). REPHQA: Evaluating Readability of large Language Models in Public Health Question Answering. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2509.16360
- Schoeninger, G. (2025, February 11). *Why GRPO is Important and How it Works*. Oxen.ai. https://ghost.oxen.ai/why-grpo-is-important-and-how-it-works/
- Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *arXiv.org*. https://arxiv.org/abs/2305.18290
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*. https://doi.org/10.48550/arXiv.1707.06347
- Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., & Guo, D. (2024). DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300v3*. https://doi.org/10.48550/arXiv.2402.03300
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, *30*, 5998 – 6008. https://arxiv.org/abs/1706.03762
- Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., Liu, X., Lin, H., Lin, Z., Ma, B., Sheng, G., Tong, Y., Zhang, C., Zhang, W., Zhu, H., … Wang, M. (2025). DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*. https://doi.org/10.48550/arXiv.2503.14476

References

- Zhang, Y. (2025, February 7). DeepSeek-R1 Dissection: Understanding PPO & GRPO Without Any Prior Reinforcement Learning Knowledge. *Hugging Face*. https://huggingface.co/blog/NormalUhr/grpo