Stabilizing Reasoning in Medical LLMs with Continued Pretraining and Reasoning Preference Optimization

Wataru Kawakami*
Graduate School of Information Science and Technology
The University of Tokyo
Tokyo, Japan
wataru.kawakami.kw@gmail.com

Keita Suzuki
Preferred Networks Inc.
Tokyo, Japan
keitasuzuki@preferred.jp

Junichiro Iwasawa
Preferred Networks Inc.
Tokyo, Japan
iwasawa@preferred.jp

Abstract

Large Language Models (LLMs) show potential in medicine, yet clinical adoption is hindered by concerns over factual accuracy, language-specific limitations (e.g., Japanese), and critically, their reliability when required to generate reasoning explanations—a prerequisite for trust. This paper introduces Preferred-MedLLM-Qwen-72B, a 72B-parameter model optimized for the Japanese medical domain to achieve both high accuracy and stable reasoning. We employ a two-stage fine-tuning process on the Qwen2.5-72B base model: first, Continued Pretraining (CPT) on a comprehensive Japanese medical corpus instills deep domain knowledge. Second, Reasoning Preference Optimization (RPO), a preference-based method, enhances the generation of reliable reasoning pathways while preserving high answer accuracy. Evaluations on the Japanese Medical Licensing Exam benchmark (IgakuQA) show Preferred-MedLLM-Qwen-72B achieves state-of-the-art performance (0.868 accuracy), surpassing strong proprietary models like GPT-4o (0.866). Crucially, unlike baseline or CPT-only models which exhibit significant accuracy degradation (up to 11.5% and 3.8% respectively on IgakuQA) when prompted for explanations, our model maintains its high accuracy (0.868) under such conditions. This highlights RPO's effectiveness in stabilizing reasoning generation. This work underscores the importance of optimizing for reliable explanations alongside accuracy. We release the Preferred-MedLLM-Qwen-72B model weights to foster research into trustworthy LLMs for specialized, high-stakes applications.

1 Introduction

The rapid advancement of Large Language Models (LLMs) marks a significant milestone in artificial intelligence, showcasing impressive abilities in understanding, synthesizing, and reasoning with complex information across various domains [OpenAI et al., 2024, Grattafiori et al., 2024, Qwen Team et al., 2024, Preferred Elements et al., 2024]. In medicine, these models hold immense potential, capable of achieving expert-level performance on benchmarks [Nori et al., 2024, Singhal et al., 2025]

*This work has been done when W.K. worked as a part-time engineer at Preferred Networks.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

and offering utility in applications such as clinical text summarization [Veen et al., 2024], patient interaction [Cosentino et al., 2024], diagnostic decision support [Tu et al., 2025] and information extraction [Mou et al., 2024]. Foundation models, especially those developed for English, have pushed performance boundaries in medical domains [Saab et al., 2024, Brodeur et al., 2024].

However, the deployment of LLMs in high-stakes clinical settings faces substantial obstacles. Factual inaccuracies (hallucinations) persist [Kim et al., 2025], critically undermining the trust required for medical applications. Performance also varies significantly across languages [Xie et al., 2024], necessitating specialized models for contexts like Japan's healthcare system. A crucial, yet often overlooked, challenge is the observed performance degradation when complex reasoning is required to solve the task [Chen et al., 2024], especially in multilingual contexts [Xie et al., 2024]. Since clinicians often need to understand the model's rationale for verification and trust, this instability presents a major barrier to adoption.

Domain-specific adaptation is a common strategy for building LLMs for the medical domain. Continued Pretraining (CPT) on specialized corpora effectively infuses domain knowledge [Gururangan et al., 2020, Chen et al., 2023, Christophe et al., 2024], while instruction fine-tuning aligns models to specific tasks or styles [Singhal et al., 2025, Zhang et al., 2023]. While efforts exist for Japanese medical LLMs [Sukeda et al., 2024, Sukeda, 2024], they often use smaller models or focus on instruction following, potentially lacking deep knowledge integration on large models [Gekhman et al., 2024] or explicit optimization for reasoning stability.

To address the dual need for high accuracy and reliable reasoning explanations in the Japanese medical domain, we introduce Preferred-MedLLM-Qwen-72B. We propose a two-stage fine-tuning process applied to the Qwen2.5-72B foundation model [Qwen Team et al., 2024]. The first stage involves CPT using a comprehensive Japanese medical corpus, including the Japanese national licensing exam materials to embed deep domain knowledge. The second stage employs Reasoning Preference Optimization (RPO) [Pang et al., 2024], a preference optimization technique extending Direct Preference Optimization (DPO) [Rafailov et al., 2023], using a curated dataset comparing ground-truth and model-generated explanations. This RPO stage specifically aims to enhance the model's ability to generate stable, high-quality reasoning pathways while maintaining accuracy.

Our core contributions are:

- We propose and evaluate a two-stage CPT+RPO fine-tuning approach designed to instill deep domain knowledge while specifically addressing the critical issue of performance instability when LLMs generate reasoning explanations in specialized domains.
- We demonstrate through evaluation on the Japanese Medical Licensing Exam benchmark (IgakuQA) [Kasai et al., 2023] that our resulting model, Preferred-MedLLM-Qwen-72B, achieves state-of-the-art accuracy (surpassing GPT-4o) and, crucially, maintains this high accuracy when required to provide explanations, validating the effectiveness of RPO for reasoning stabilization where CPT-only models falter.
- We release the Preferred-MedLLM-Qwen-72B model weights, providing a more reliable LLM for the Japanese medical domain and facilitating further research into trustworthy specialized LLMs.

This work underscores the importance of optimizing not just for accuracy but also for the reliability and transparency of the reasoning process, presenting a methodology to build more dependable LLMs for high-stakes, specialized domains and non-English languages.

2 Related Works

The application of Large Language Models (LLMs) in medicine is rapidly evolving. Foundation models, such as OpenAI's GPT series [Nori et al., 2023a], o1 series [Xie et al., 2024, Brodeur et al., 2024] and Google's Gemini family [Saab et al., 2024], have demonstrated remarkable reasoning capabilities on various English medical benchmarks, often achieving high performance with minimal prompting [Nori et al., 2024]. However, significant challenges remain, including factual inaccuracies or hallucinations, and inconsistent performance across different languages and cultural contexts [Xie et al., 2024]. These limitations hinder reliable deployment in high-stakes clinical settings and motivate the need for domain- and language-specific adaptations.

A primary strategy for adapting LLMs is further training on specialized data. Two common approaches are Continued Pretraining (CPT) and instruction fine-tuning. CPT involves extending the pretraining phase on large, domain-specific corpora. This approach, discussed conceptually by Gururangan et al. [2020] and implemented in models like Meditron-70B [Chen et al., 2023] and others [Christophe et al., 2024], is effective for infusing deep domain knowledge into the model, which is considered a more effective method for knowledge addition compared to instruction tuning [Gekhman et al., 2024]. Instruction fine-tuning, conversely, aligns models to follow specific instructions or interaction styles, as seen in models like MedPaLM for English medical QA [Singhal et al., 2025], HuatuoGPT for Chinese consultations [Zhang et al., 2023].

Work has also emerged specifically within the Japanese medical domain. For instance, Sukeda [2024] developed LLMs by applying instruction tuning to Japanese medical data. While valuable, such efforts have often utilized smaller base models (e.g., 7B parameters) or primarily focused on instruction following. This leaves a potential gap regarding the deep knowledge integration achievable via CPT on larger foundation models and the explicit optimization of reasoning processes, a critical step for ensuring reliability in real-world clinical and research applications.

Beyond fine-tuning model weights, prompt engineering techniques like Medprompt [Nori et al., 2023b] aim to elicit better medical performance from generalist models at inference time. However, the effectiveness and complexity of such strategies can vary [Nori et al., 2024], potentially favoring methods that embed specialized knowledge and reasoning capabilities directly into the model parameters through training. Enhancing the inherent reasoning capabilities of LLMs is also an active research area, exploring techniques such as reinforcement learning [Xie et al., 2024, DeepSeek-AI et al., 2025], reasoning with Reinforced Fine-Tuning (ReFT) [Luong et al., 2024], and preference optimization methods like Direct Preference Optimization (DPO) [Rafailov et al., 2023] and its variants [Pang et al., 2024].

A critical gap, particularly relevant for clinical trust and adoption, is the performance degradation when models are required to generate step-by-step reasoning explanations. Our work addresses this specific challenge. We combine deep domain adaptation via CPT on a comprehensive Japanese medical corpus with Reasoning Preference Optimization (RPO) [Pang et al., 2024], a DPO variant. RPO is specifically chosen here for its focus on not only learning preferences, but also on stabilizing the likelihood of the preferred reasoning path, aligning with our goal of generating reliable explanations alongside accurate answers. Unlike generalist models evaluated on medicine, or domain-specific models fine-tuned via CPT or instruction following, our approach explicitly targets the Japanese medical context and optimizes for both high accuracy and stable, high-quality reasoning explanations. We demonstrate that this combination achieves state-of-the-art performance on the complex Japanese Medical Licensing Exam (IgakuQA), surpassing strong proprietary models, and crucially, maintains this performance even when explicitly prompted for explanations, overcoming the instability observed in models without RPO.

3 Methods

Our approach involves a two-stage fine-tuning process applied to a powerful base LLM to create Preferred-MedLLM-Qwen-72B, specifically tailored for the Japanese medical domain with enhanced reasoning capabilities.

3.1 Stage 1: Continued Pretraining (CPT) for Domain Adaptation

We selected Qwen2.5-72B [Qwen Team et al., 2024] as the base model. This choice was based on its state-of-the-art performance on general benchmarks at the time of experimentation, its strong multilingual capabilities, which provide a solid base for Japanese language understanding.

The primary objective of the first stage was to infuse the base Qwen2.5-72B with comprehensive, up-to-date Japanese medical knowledge, addressing potential gaps not covered by its general pretraining corpus. In contrast to instruction tuning, which mainly adjusts response style based on existing knowledge [Ouyang et al., 2022], CPT enables the model to acquire and integrate new domain-specific information [Gekhman et al., 2024].

To achieve this, we utilized an original medical corpus, a collection of Japanese medical texts. This dataset includes explanations for past problems from the Japanese Medical Licensing Exam (JMLE) up to the year 2017. mr Crucially, to prevent data leakage into our evaluation set, this corpus only includes JMLE materials published up to the year 2017, ensuring no overlap with the IgakuQA benchmark which covers 2018–2022 [Kasai et al., 2023].

Training a 72B parameter model presents significant computational challenges. We addressed this by employing QLoRA (Quantized Low-Rank Adaptation) [Dettmers et al., 2023]. Here, we combined 4-bit NormalFloat (NF4) quantization [Dettmers and Zettlemoyer, 2022] with Low-Rank Adaptation (LoRA) [Hu et al., 2021]. The CPT phase itself was conducted for two epochs. We employed the AdamW optimizer with cosine scheduling with warmup. This stage of training was carried out on an internal compute cluster, utilizing four NVIDIA A100 80GB GPUs.

3.2 Stage 2: Reasoning Preference Optimization (RPO)

Following CPT, the second stage focused on Reasoning Preference Optimization (RPO). The main goal here was to enhance the model's ability to generate not only accurate answers but also coherent and reliable reasoning explanations. This specifically aimed to address the performance instability observed in the CPT-only model when it was prompted to provide explanations alongside its answers.

To facilitate RPO, we curated a specialized preference dataset. This dataset was built using JMLE problems sourced up to 2017, ensuring distinction from the data used in IgakuQA. The curation process involved several steps. First, we prompted the CPT model (from Section 3.1) using a three-shot template ("Question: {JMLE Question}\n{Options}\n Explanation:{Explanation for the question}\n Answer:{Answer}"]) to generate a response containing both a step-by-step explanation and a final conclusion answering the question. Second, these generated responses were automatically categorized based on whether their final answer matched the ground truth correct answer for the respective JMLE problem. Responses with the correct final answer were labeled as "Chosen Responses" while those with incorrect final answers were labeled as "Rejected Responses". Third, we prepared "Ground Truth Responses", which consisted of the correct final answer paired with a verified explanation. It should be noted that these ground truth responses with verified explanations were also included in the data for continued pretraining.

With the preference data prepared, we employed Reasoning Preference Optimization (RPO) [Pang et al., 2024], a variant of DPO [Rafailov et al., 2023], implemented using the Hugging Face TRL library. RPO learns directly from preference pairs. We established a preference hierarchy for each JMLE problem: Ground Truth Response $>$ Chosen Response $>$ Rejected Response. This hierarchy guides the RPO loss computation, which consists of a weighted negative log-likelihood (NLL) loss on the chosen preferences together with the DPO loss, by creating preference pairs like (Ground Truth, Chosen), (Ground Truth, Rejected), and (Chosen, Rejected). This structure explicitly trains the model to favor responses demonstrating correctness and high-quality reasoning (represented by the ground truth preference) over those with correct answers but potentially weaker generated reasoning, while strongly penalizing incorrect answers. The underlying assumption is that the curated ground truth explanations embody a higher standard of reasoning compared to the model's intermediate generated explanations, even if the latter lead to a correct answer.

The RPO training phase was applied to the CPT model weights. For computational efficiency, we again utilized QLoRA. The model underwent training for one epoch over the entire preference dataset. We set the RPO alpha parameter, which weights the NLL loss relative to the DPO loss, to 10. This RPO stage was performed using two A100 GPUs.

4 Results

This section details the performance evaluation of Preferred-MedLLM-Qwen-72B. We first present its results on the primary benchmark, the Japanese National Medical Licensing Exam (IgakuQA), comparing it against baseline and proprietary models under different prompting conditions. We then provide an ablation study to dissect the contributions of the Continued Pretraining (CPT) and Reasoning Preference Optimization (RPO) stages. Finally, we assess the model's generalization capabilities on other relevant Japanese medical question-answering datasets.

Table 1: Performance comparison on the IgakuQA benchmark. Scores represent the average scores over the 2018–2022 exams, normalized by the total score. Evaluations were conducted under standard 3-shot prompting and 3-shot prompting explicitly requiring explanations ('w/ explanation').
*Scores for GPT-4 and GPT-3.5 were taken from Kasai et al. [2023].

Model	IgakuQA (3-shot)	IgakuQA (3-shot w/ explanation)
Preferred-MedLLM-Qwen-72B	**0.868**	0.868
GPT-4o	0.866	**0.881**
GPT-4 Turbo	0.812	0.814
Qwen2.5-72B-Instruct	0.802	0.822
Qwen2.5-72B	0.802	0.710
Llama3-Preferred-MedSwallow-70B	0.795	0.744
Llama-3.3-Swallow-70B-v0.4	0.787	0.755
GPT-4*	0.782	-
GPT-4o mini	0.751	0.759
Llama-3-Swallow-70B-v0.1	0.701	0.637
sarashina2-70b	0.561	0.549
GPT-3.5*	0.550	-

4.1 Performance on the IgakuQA Benchmark

We evaluated our model using the IgakuQA benchmark [Kasai et al., 2023], which comprises questions from the Japanese National Medical Licensing Exam spanning the years 2018 to 2022. Performance is reported as the average score achieved across these five years, calculated as the total points earned divided by the total possible points to account for varying question weights.

Evaluations were conducted under two distinct prompting scenarios. The first employed a standard 3-shot setting, where the prompt included three examples consisting of a question, options, and the correct answer, preceding the target question. The specific examples used were identical to those in the original IgakuQA implementation. As shown in Table 1, Preferred-MedLLM-Qwen-72B achieved a score of 0.868 in this standard setting. This performance surpasses that of the base Qwen2.5-72B (0.802) and the instruction-tuned Qwen2.5-72B-Instruct (0.802). It also slightly exceeds the score obtained by GPT-4o (0.866) and is considerably higher than GPT-4-Turbo (0.812) and the previously reported GPT-4 score (0.782) [Kasai et al., 2023].

The second evaluation scenario, termed "3-shot w/ explanation", was designed specifically to assess performance stability when the model is explicitly required to generate its reasoning process. In this setting, the prompt instructed the model to provide a step-by-step explanation before delivering the final answer (similar in structure to the examples shown in Table **??**). Under this condition, Preferred-MedLLM-Qwen-72B maintained its high accuracy, achieving an identical score of 0.868. This stability contrasts with the behavior of the baseline Qwen2.5-72B model, whose performance decreased substantially from 0.802 in the standard setting to 0.710 when prompted for explanations. While GPT-4o demonstrated strong performance in this setting (0.881), our model outperformed the instruction-tuned Qwen2.5-72B-Instruct (0.822). These results underscore the effectiveness of our combined CPT and RPO methodology in yielding a model that not only achieves high accuracy but also preserves this accuracy when generating explanatory reasoning, a key desideratum for clinical utility.

4.2 Ablation Study: Impact of CPT and RPO

To delineate the individual and combined contributions of CPT and RPO, we conducted an ablation study using the IgakuQA benchmark. The results are presented in Table 2. We also included a comparison using standard DPO instead of RPO.

The baseline Qwen2.5-72B model achieved scores of 0.802 (3-shot) and 0.710 (3-shot w/ explanation). Applying only CPT (Stage 1) significantly improved performance in the standard setting to 0.867, confirming the effectiveness of CPT for domain knowledge infusion. However, this CPT-only model exhibited a performance drop to 0.834 when explanations were required, highlighting the reasoning

Table 2: Ablation study on IgakuQA showing the impact of CPT, RPO, and DPO applied to Qwen2.5-72B. Scores represent average accuracy under standard 3-shot and 3-shot with explanation ('w/ explanation') settings.

Model	IgakuQA (3-shot)	IgakuQA (3-shot w/ explanation)
Preferred-MedLLM-Qwen-72B (CPT+RPO)	**0.868**	**0.868**
Qwen2.5-72B + CPT + DPO	**0.868**	0.848
Qwen2.5-72B + CPT	0.867	0.834
Qwen2.5-72B + RPO	0.838	0.807
Qwen2.5-72B	0.802	0.710

instability issue motivating our work. Conversely, applying RPO directly to the base Qwen2.5-72B model resulted in scores of 0.838 (3-shot) and 0.807 (3-shot w/ explanation). This indicates that RPO alone can enhance both accuracy and reasoning stability compared to the baseline, likely by better aligning the model's inherent reasoning pathways, although it lacks the deep domain knowledge provided by CPT.

Combining CPT with standard DPO yielded scores of 0.868 (3-shot) and 0.848 (3-shot w/ explanation). While achieving high accuracy comparable to our final model in the standard setting, this configuration still showed a noticeable decrease in performance when generating explanations.

Finally, our proposed two-stage approach, combining CPT with RPO (Preferred-MedLLM-Qwen-72B), achieved a score of 0.868 in the standard 3-shot setting and, crucially, maintained this exact score of 0.868 in the 3-shot setting requiring explanations. This demonstrates perfect stability under explanation generation for this benchmark. These ablation results provide strong evidence for the synergy between CPT and RPO. CPT is essential for incorporating domain-specific knowledge, while RPO effectively refines the model's reasoning generation, ensuring consistent accuracy even when explanations are explicitly prompted. Furthermore, the comparison between CPT+DPO and CPT+RPO suggests that RPO, potentially due to its inclusion of an NLL loss component alongside the DPO loss, offers superior stabilization for the reasoning process in this context.

4.3 Performance on Other Japanese Medical Benchmarks

To evaluate the generalizability of the improvements imparted by our fine-tuning methodology, we assessed Preferred-MedLLM-Qwen-72B on a selection of other Japanese medical question-answering benchmarks. These evaluations were performed in a zero-shot setting to probe the model's inherent capabilities without task-specific examples. The benchmarks included Japanese translations of MedQA [Jin et al., 2020], MedMCQA [Pal et al., 2022], PubMedQA [Jin et al., 2019] (using translations from Jiang et al. [2024]), and the medicine related subset ("anatomy", "clinical_knowledge", "college_medicine", "medical_genetics", "professional_medicine") of MMMLU [OpenAI, 2024, Hendrycks et al., 2020].

The results, summarized in Table 3, indicate that Preferred-MedLLM-Qwen-72B generally outperforms both the base Qwen2.5-72B and the Qwen2.5-72B-Instruct models across these diverse tasks. Our model achieved the highest scores on the Japanese medical related tasks of MMMLU, and the Japanese translated versions of MedQA, and MedMCQA, resulting in the highest average score (0.716) among the compared models. Furthermore, comparing our final Preferred-MedLLM-Qwen-72B model (CPT+RPO) to the model after only the CPT stage (Qwen2.5-72B + CPT in Table 3), we observe that the addition of RPO yields further small but consistent improvements across most of these diverse benchmarks, increasing the average score slightly from 0.710 to 0.716. This consistent performance advantage suggests that the benefits of the combined CPT and RPO stages, including the potential positive impact of RPO on generalizability, extend effectively beyond the specific format of the Japanese Medical Licensing Exam to broader medical question-answering scenarios in Japanese.

5 Conclusion and Discussions

In this work, we introduced Preferred-MedLLM-Qwen-72B, a large language model specifically fine-tuned for the Japanese medical domain. Our primary goal was to develop an LLM exhibiting not

Table 3: Zero-shot performance on various Japanese medical QA benchmarks. Scores represent accuracy on Japanese translations of MedQA, MedMCQA, PubMedQA, and relevant subsets ("anatomy", "clinical_knowledge", "college_medicine", "medical_genetics", "professional_medicine") of MMMLU. Benchmark translations were sourced from Jiang et al. [2024].

Model (0-shot)	MMMLU (jp, med)	MedQA (jp)	MedMCQA (jp)	PubMedQA (jp)	Average
Preferred-MedLLM-Qwen-72B	**0.800**	**0.684**	**0.602**	**0.779**	**0.716**
Qwen2.5-72B + CPT	0.799	0.669	0.598	0.775	0.710
Qwen2.5-72B	0.797	0.652	0.591	**0.779**	0.705
Qwen2.5-72B-Instruct	0.795	0.626	0.601	0.714	0.684

only high accuracy but also reliable performance when generating reasoning explanations crucial for clinical trust. We proposed a two-stage fine-tuning process combining Continued Pretraining (CPT) for deep domain knowledge infusion and Reasoning Preference Optimization (RPO) for enhancing reasoning alignment and stability. Our core contribution lies in demonstrating that applying RPO subsequent to CPT effectively mitigates the reasoning instability—a performance degradation observed when explicit explanations are required—that can manifest in models trained with CPT alone or even CPT combined with standard DPO.

Our empirical results validate this approach. Preferred-MedLLM-Qwen-72B achieves state-of-the-art performance on the challenging Japanese Medical Licensing Exam benchmark (IgakuQA), surpassing its base model and competitive proprietary models like GPT-4o in standard evaluations. More significantly, ablation studies confirmed that while CPT is vital for knowledge and accuracy gains, the RPO stage is crucial for maintaining this high performance when the model is prompted to generate step-by-step explanations. This stability under explanation is a key differentiator from baseline and CPT-only models. Furthermore, the model demonstrated strong performance across other Japanese medical QA benchmarks, suggesting the benefits of our CPT+RPO methodology generalize beyond the specific JMLE task format. This work presents a promising methodology for building more reliable and transparent AI systems by strategically optimizing for both accuracy and stable reasoning generation.

While these findings are encouraging, we acknowledge several limitations and identify avenues for future investigation. First, our evaluation primarily focused on multiple-choice question-answering benchmarks. Assessing performance on a broader spectrum of clinical tasks, such as medical report generation [Kanithi et al., 2024], summarization [Veen et al., 2024], and dialogue systems [Johri et al., 2025], is essential to fully understand the model's capabilities. Second, the preference data used for RPO was generated semi-automatically based on ground truth answers and model outputs. Future work could explore incorporating expert human feedback into the preference framework, which might further enhance reasoning quality, although scalability must be considered. Third, we did not investigate combining RPO with instruction tuning; exploring this combination could reveal complementary benefits for reasoning stabilization and overall performance. Fourth, the significant computational resources required to train and deploy 72B parameter models pose a practical challenge; future research should explore applying our CPT+RPO methodology to smaller, more efficient architectures or investigate model distillation techniques. Finally, rigorous real-world clinical validation remains a critical next step to evaluate the practical utility, safety, and trustworthiness of the model and its explanations within actual healthcare settings.

In conclusion, Preferred-MedLLM-Qwen-72B represents a significant step towards developing specialized LLMs suitable for demanding domains like Japanese medicine. By strategically combining CPT for knowledge and RPO for reasoning alignment, we have created a model that achieves high accuracy while crucially maintaining performance stability during explanation generation. This highlights an effective pathway for building more reliable and transparent AI systems, paving the way for their responsible integration into clinical practice and other high-stakes applications.

Software and Data

The trained model can be accessed at `https://huggingface.co/pfnet/Preferred-MedLLM-Qwen-72B`.

Acknowledgements

We would like to thank the Preferred Networks cluster team members for the infrastructure support.

References

Peter G. Brodeur, Thomas A. Buckley, Zahir Kanjee, Ethan Goh, Evelyn Bin Ling, Priyank Jain, Stephanie Cabral, Raja-Elie Abdulnour, Adrian Haimovich, Jason A. Freed, Andrew Olson, Daniel J. Morgan, Jason Hom, Robert Gallo, Eric Horvitz, Jonathan Chen, Arjun K. Manrai, and Adam Rodman. Superhuman performance of a large language model on the reasoning tasks of a physician. *arXiv preprint arXiv:2412.10849*, 2024. URL `https://arxiv.org/abs/2412.10849`.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. HuatuoGPT-o1, towards medical complex reasoning with LLMs. *arXiv preprint arXiv:2412.18925*, 2024. URL `https://arxiv.org/abs/2412.18925`.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. MEDITRON-70B: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023. URL `https://arxiv.org/abs/2311.16079`.

Clement Christophe, Tathagata Raha, Svetlana Maslenkova, Muhammad Umar Salman, Praveenkumar Kanithi, Marco AF Pimentel, and Shadab Khan. Beyond fine-tuning: Unleashing the potential of continuous pretraining for clinical LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10549–10561, Miami, Florida, USA, 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.findings-emnlp.618/`.

Justin Cosentino, Anastasiya Belyaeva, Xin Liu, Nicholas A. Furlotte, Zhun Yang, Chace Lee, Erik Schenck, Yojan Patel, Jian Cui, Logan Douglas Schneider, Robby Bryant, Ryan G. Gomes, Allen Jiang, Roy Lee, Yun Liu, Javier Perez, Jameson K. Rogers, Cathy Speed, Shyam Tailor, Megan Walker, Jeffrey Yu, Tim Althoff, Conor Heneghan, John Hernandez, Mark Malhotra, Leor Stern, Yossi Matias, Greg S. Corrado, Shwetak Patel, Shravya Shetty, Jiening Zhan, Shruthi Prabhakara, Daniel McDuff, and Cory Y. McLean. Towards a personal health large language model. *arXiv preprint arXiv:2406.06474*, 2024. URL `https://arxiv.org/abs/2406.06474`.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun

Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL `https://arxiv.org/abs/2501.12948`.

Tim Dettmers and Luke Zettlemoyer. The case for 4-bit precision: k-bit inference scaling laws. *arXiv preprint arXiv:2212.09720*, 2022. URL `https://arxiv.org/abs/2212.09720`.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36:10088–10115, 2023. URL `https://dl.acm.org/doi/10.5555/3666122.3666563`.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning LLMs on new knowledge encourage hallucinations? *arXiv preprint arXiv:2405.05904*, 2024. URL `https://arxiv.org/abs/2405.05904`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, and Tobias Speckbacher. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL `https://arxiv.org/abs/2407.21783`.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, 2020. Association for Computational Linguistics. URL `https://aclanthology.org/2020.acl-main.740/`.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. URL `https://arxiv.org/abs/2009.03300`.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. URL `https://arxiv.org/abs/2106.09685`.

Junfeng Jiang, Jiahao Huang, and Akiko Aizawa. JMedBench: A benchmark for evaluating japanese biomedical large language models. *arXiv preprint arXiv:2409.13317*, 2024. URL `https://arxiv.org/abs/2409.13317`.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020. URL `https://arxiv.org/abs/2009.13081`.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China, 2019. Association for Computational Linguistics. URL `https://aclanthology.org/D19-1259/`.

Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I. Schlessinger, Shannon Wongvibulsin, Leandra A. Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M. Van Allen, David Kim, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature Medicine*, pages 1–10, 2025. URL `https://www.nature.com/articles/s41591-024-03328-5`.

Praveen K. Kanithi, Clément Christophe, Marco AF Pimentel, Tathagata Raha, Nada Saadi, Hamza Javed, Svetlana Maslenkova, Nasir Hayat, Ronnie Rajan, and Shadab Khan. MEDIC: Towards a comprehensive framework for evaluating LLMs in clinical applications. *arXiv preprint arXiv:2409.07314*, 2024. URL `https://arxiv.org/abs/2409.07314`.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. Evaluating GPT-4 and ChatGPT on japanese medical licensing examinations, 2023. URL `https://arxiv.org/abs/2303.18027`.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, Lizhou Fan, Eugene Park, Tristan Lin, Joonsik Yoon, Wonjin Yoon, Maarten Sap, Yulia Tsvetkov, Paul Liang, Xuhai Xu, Xin Liu, Daniel McDuff, Hyeonhoon Lee, Hae Won Park, Samir Tulebaev, and Cynthia Breazeal. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*, 2025. URL `https://arxiv.org/abs/2503.05777`.

Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. ReFT: Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967*, 2024. URL `https://arxiv.org/abs/2401.08967`.

Yongli Mou, Hanbin Chen, Gwendolyn Isabella Lode, Daniel Truhn, Sulayman Sowe, and Stefan Decker. RadLink: Linking clinical entities from radiology reports. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 443–449. IEEE, 2024. URL `https://ieeexplore.ieee.org/document/10852450`.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023a. URL `https://arxiv.org/abs/2311.16452`.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023b. URL `https://arxiv.org/abs/2311.16452`.

Harsha Nori, Naoto Usuyama, Nicholas King, Scott Mayer McKinney, Xavier Fernandes, Sheng Zhang, and Eric Horvitz. From Medprompt to o1: Exploration of run-time strategies for medical challenge problems and beyond. *arXiv preprint arXiv:2411.03590*, 2024. URL `https://arxiv.org/abs/2411.03590`.

OpenAI. openai/MMMLU, 2024. `https://huggingface.co/datasets/openai/MMMLU`.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel

Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, and Jong Wook Kim. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. URL `https://arxiv.org/abs/2410.21276`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. URL `https://dl.acm.org/doi/10.5555/3600270.3602281`.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. *arXiv preprint arXiv:2203.14371*, 2022. URL `https://arxiv.org/abs/2203.14371`.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. *arXiv preprint arXiv:2404.19733*, 2024. URL `https://arxiv.org/abs/2404.19733`.

Preferred Elements, Kenshin Abe, Kaizaburo Chubachi, Yasuhiro Fujita, Yuta Hirokawa, Kentaro Imajo, Toshiki Kataoka, Hiroyoshi Komatsu, Hiroaki Mikami, Tsuguo Mogami, Shogo Murai, Kosuke Nakago, Daisuke Nishino, Toru Ogawa, Daisuke Okanohara, Yoshihiko Ozaki, Shotaro Sano, Shuji Suzuki, Tianqi Xu, and Toshihiko Yanase. PLaMo-100B: A ground-up language model designed for japanese proficiency. *arXiv preprint arXiv:2410.07563*, 2024. URL `https://arxiv.org/abs/2410.07563`.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL `https://arxiv.org/abs/2412.15115`.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023. URL `https://arxiv.org/abs/2305.18290`.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024. URL `https://arxiv.org/abs/2404.18416`.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8, 2025. URL `https://www.nature.com/articles/s41591-024-03423-7`.

Issey Sukeda. Development and bilingual evaluation of japanese medical large language model within reasonably low computational resources. *arXiv preprint arXiv:2409.11783*, 2024. URL `https://arxiv.org/abs/2409.11783`.

Issey Sukeda, Risa Kishikawa, and Satoshi Kodera. 70B-parameter large language models in japanese medical question-answering. *arXiv preprint arXiv:2406.14882*, 2024. URL `https://arxiv.org/abs/2406.14882`.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. Towards conversational diagnostic artificial intelligence. *Nature*, pages 1–9, 2025. URL `https://www.nature.com/articles/s41586-025-08866-7`.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142, 2024. URL `https://www.nature.com/articles/s41591-024-02855-5`.

Yunfei Xie, Juncheng Wu, Haoqin Tu, Siwei Yang, Bingchen Zhao, Yongshuo Zong, Qiao Jin, Cihang Xie, and Yuyin Zhou. A preliminary study of o1 in medicine: Are we closer to an AI doctor? *arXiv preprint arXiv:2409.15277*, 2024. URL `https://arxiv.org/abs/2409.15277`.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. HuatuoGPT, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*, 2023. URL `https://arxiv.org/abs/2305.15075`.