Tiny Reward Models

Sarah Pan [1]

Abstract

Large decoder-based language models have become the dominant architecture for reward modeling in reinforcement learning from human feedback (RLHF). However, as reward models are increasingly deployed in test-time strategies, their inference costs become a growing concern. We present TinyRM, a family of small, bidirectional masked language models (MLMs) with as few as 400 million parameters, that rival the capabilities of models over 175 times larger on reasoning and safety preference modeling tasks. TinyRM combines FLAN-style prompting, Directional Low-Rank Adaptation (DoRA), and layer freezing to achieve strong performance on RewardBench, despite using significantly fewer resources. Our experiments suggest that small models benefit from domain-specific tuning strategies, particularly in reasoning, where lightweight finetuning methods are especially effective. While challenges remain in building generalist models and conversational preference modeling, our preliminary results highlight the promise of lightweight bidirectional architectures as efficient, scalable alternatives for preference modeling.

1. Introduction

Reinforcement learning from human feedback (RLHF) has become a foundational approach for aligning large language models (LLMs) with human preferences (Christiano et al., 2023; Ouyang et al., 2022). However, the success of RLHF hinges entirely on the quality of the reward model (RM) in evaluation (Shen et al., 2023). In any RM-reliant setting, the strength of signal provided is important as optimizing against an inaccurate reward model limits overall effectiveness (Gao et al., 2022).

[1]Massachusetts Institute of Technology and answer.ai, Cambridge, USA. Correspondence to: Sarah Pan <sarahpan@mit.edu>.

Published at ICML 2025 Workshop on Efficient Systems for Foundation Models, Vancouver, Canada.

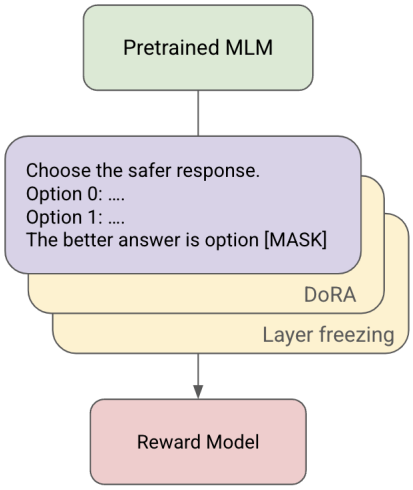


Figure 1. Overview of our finetuning pipeline: DoRA and layer freezing combined with FLAN-style, cloze prompting convert a pretrained MLM into a reward model.

Recent approaches to training strong reward models rely on scaling up decoder-based LLMs as RMs, under the assumption that bigger models generalize better (Nvidia et al., 2024; Wang et al., 2024; Winata et al., 2025). However, while large RMs represent a one-time expense in train-time pipelines such as RLHF, their presence in new paradigms such as agentic workflow routing, synthetic data filtering, and inference-time process supervision introduces recurrent costs (Gunasekar et al., 2023; Lu et al., 2023; Luo et al., 2024).

While multi-billion parameter RMs achieve strong performance, new RM-based test-time decoding strategies present substantial compute and memory overhead. This is especially unjustified if it is unclear whether preference modeling benefits from the same scaling laws as one-shot next token generation (Hou et al., 2024; Chen et al., 2025; Song et al., 2023). This motivates our work into training efficient RMs.

In this work, we propose TinyRM, a family of lightweight, bidirectional masked language models (MLMs) that not only perform competitively on reward modeling benchmarks but

also provide insight into eliciting strong language *capabilities* from models with strong language *understanding*. We combine FLAN-style prompting (Wei et al., 2022), Directional Low-Rank Adaptation (DoRA) (Liu et al., 2024b), and layer freezing to create efficient specialists capable of high-quality preference modeling across different domains.

Our contributions to this end are threefold:

1. We demonstrate that bidirectional MLMs can serve as effective RMs, rivaling models over 175x larger in certain tasks.

2. We introduce and evaluate a combination of efficient finetuning techniques—specifically FLAN-style prompting, Directional Low-Rank Adaptation (DoRA), and layer freezing—for adapting MLMs to domain-specific preference modeling.

3. We provide empirical insights into how these techniques interact across task domains, revealing the surprising strength of small-scale models on reasoning and safety tasks, while highlighting limitations in open-ended conversational tasks.

2. Background & Related Work

2.1. Reward Modeling

Reward models are a cornerstone of reinforcement learning from human feedback (RLHF), which steers generative models toward human-aligned behaviors (Christiano et al., 2023). Contemporary RMs are typically instantiated from autoregressive large language models (LLMs), whose parameter counts have increased significantly since their inception (Ouyang et al., 2022).

However, recent work suggests that RMs have significant potential outside of RLHF. For instance, they play key roles in routing agentic workflow, data filtration, and process supervision during inference (Gunasekar et al., 2023; Lu et al., 2023; Luo et al., 2024).

In many of these new RM application settings, the efficiency of the reward model presents a new challenge. For instance, in guided decoding pipelines such as that of Chaffin et al. (2022), inference of the discriminator model scales with the number of potential completions explored. This is different from the RLHF setting, where performing inference on a large reward model represents a one-time cost which is eventually "amortized" by repeated downstream use of the finetuned model.

RewardBench (Lambert et al., 2024) evaluates reward models on four categories: Chat, Chat-Hard, Reasoning, and Safety. In our work, we merge Chat and Chat-Hard into a single 'Chat' category for simplicity. This domain covers open-ended conversational preferences. Reasoning covers math and coding tasks while Safety evaluates refusal capabilities for harmful prompts.

2.2. Encoders are Strong Language Understanders

Prior work suggests that encoder and encoder-decoder language models, offer an efficient foundation for tasks requiring natural language understanding (NLU) (Devlin et al., 2019; Raffel et al., 2023; Lewis et al., 2019).

Unlike unidirectional autoregressive LLMs, encoder-based models are able to access contextual information in both directions, resulting in richer internal representations (Skean et al., 2025). This advantage has been demonstrated not just in information retrieval and NLU benchmarks, where encoders dominate the space, (Nogueira & Cho, 2020; Lewis et al., 2021) but also in language generation, where MLMs were shown to be more data efficient than decoder-based ones (Samuel, 2024).

Moreover, previous work has shown that small encoders are strong few-shot learners and can outperform decoder-based LLMs in resource constrained settings (Schick & Schütze, 2021b; Gao et al., 2021).

2.3. Small Models and Compound Objectives

Recent studies of small language models demonstrate coherent language abilities that occur despite their small size. For instance, Eldan & Li (2023) and Ghanizadeh & Dousti (2025) took data-centric approaches and demonstrated that small, decoder-based models could exhibit both creative and grammatical competence when trained on simplistic text.

Along similar lines, Steuer et al. (2023) hint that larger models may overfit to surface-level patterns while failing to fundamentally reason as evidenced by low surprisal for complex tasks.[1]

3. Experiments

For our main experiment, we trained individual models initialized from ModernBERT-Base and ModernBERT-Large (150 and 400 million parameters, respectively) (Warner et al., 2024) as Chat, Reasoning, and Safety specialists on publicly available preference data.[2] We performed a sweep across multiple hyperparameters, the specifics of which can be found in Appendix A. We also performed sweeps to train a large "All-At-Once" (AAO) model where the same finetuning framework was used with all of the domain-specific data together, at once.

[1] Surprisal has been shown to be a predictor of reading time or task difficulty for humans (Fernandez Monsalve et al., 2012).

[2] Specific datasets used can be found in Appendix B.

We leave the ablation of the specific methods used to future work. As a preliminary work, we focus on reporting empirical observations using different combinations of these strategies.

3.1. DoRA and Layer Freezing

We noticed significant performance improvements in certain domains when using a combination of Weight-Decomposed Low-Rank Adaptation (DoRA) and layer freezing. In our runs, we swept over whether DoRA was used, its LoRA rank, as well as the number of layers frozen.

DoRA is a parameter efficient finetuning method that decomposes model weights into magnitude and direction components. It then uses LoRA (Hu et al., 2021) to provide more controlled updates to the direction vector (Liu et al., 2024b). For the reasoning task, DoRA provided significant performance gains over full-rank finetuning.

In addition to DoRA, we froze lower layers of the model to preserve general language representations. This allowed for a more focused finetuning of the task-specific upper layers (Lee et al., 2019; Howard & Ruder, 2018; Yosinski et al., 2014).

3.2. Training and Evaluation Format

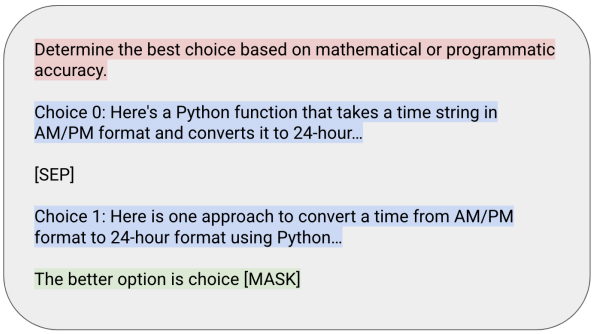
Despite extensive literature on training RMs from autoregressive LLMs, there is considerably less work on training encoder-only reward models. For this reason, early experiments consisted of testing various training schemes.

We explored three training paradigms for converting MLMs to reward models. For the standard classification approach, we applied pooling (CLS-token or mean) to hidden states with a classification head (Sun et al., 2020; Liu et al., 2024a). Despite being the conventional method for MLM adaptation, this yielded suboptimal performance.

The token-level classification approach was inspired by work with process reward models (Pan et al., 2023; Lightman et al., 2023). We assigned binary labels to tokens from chosen/rejected responses. The intuition behind this method was to derive a richer loss by imposing structure onto the output. However, this approach also yielded suboptimal performance.

Finally, we turned to instruction tuning, which has been shown to elicit stronger zero-shot capabilities on out of distribution tasks (Wei et al., 2022). For the FLAN-style masked language modeling method, we structured the task as instruction-following with masked prediction (Figure 2). This approach significantly outperformed alternatives, supporting previous results on the efficacy of instruction tuning for encoder and encoder-decoder models (Clavié et al., 2025; Chung et al., 2022; Wei et al., 2022).

Figure 2. Our FLAN-style schema consists of the problem statement, both options, and the final preference statement where the MLM makes a prediction.



Similar to that of Schick & Schütze (2021a), the FLAN-like method we employ (as shown in Figure 2) is a reformulation of the reward modeling task as a cloze question. Formally, we define the FLAN-like masked objective as minimizing cross-entropy loss over a masked token m such that $P(m|(x, y_w, y_l))$ reflects the model's preference. Here, x is the instruction prefix, y_w is the chosen completion, and y_l is the rejected one.

It is important to note that our models are trained and evaluated with visibility of both options in their contexts, which is not the case for official RewardBench evaluation.[3] Though we admit ours is not an apples-to-apples comparison, we also argue that practical deployment scenarios can be made to reflect our setup.

4. Experimental Results

Our main results are presented in Table 1. We see that the specialists initialized from ModernBERT-large perform the best of all of our experiments. The large specialist is competitive with a model 175x its size in the Reasoning task. Unlike decoder-based models, the specialist struggled the most with the Chat task. Surprisingly, the specialists initialized from ModernBERT-base perform better than the AAO model initialized from ModernBERT-large.

Chat Domain Although large specialists perform competitively for the Safety and Reasoning domains, they face unique challenges in terms of the Chat task. We hypothesize this may be attributed to the conversational finetuning performed on many open source LLMs, allowing the Chat task to be fully in-domain for those models (Grattafiori et al.,

[3]https://github.com/allenai/reward-bench

[4]We use a weighted average to consolidate the Chat and Chat-Hard categories from RewardBench into a single Chat category.

[5]At the time of submission.

Table 1. Performance (accuracy) on RewardBench. The large specialists outperform the large all-at-once model and perform comparably with a 70 billion parameter model. Our 150 million parameter specialists outperform the large 400 million parameter all-at-once model.

MODEL	CHAT[4]	REASONING	SAFETY	OVERALL
MODERNBERT-LARGE SPECIALISTS 400M	78.8	**91.2 (DORA)**	89.3	86.4
MODERNBERT-LARGE AAO 400M	71.0	76.7	79.2	75.6
MODERNBERT-BASE SPECIALISTS 150M	73.5	83.3 (DORA)	78.4	78.4
LLAMA3-STEERLM-RM 70B	89.7	90.6	**92.8**	**91.0**
OPENASSISTANT-DEBERTA-V3-LARGE-V2 400M	82.8	38.5	73.4	64.9
ANTHROPIC/CLAUDE-3-5-SONNET-20240620	**93.0**	84.7	81.6	86.4

2024). We were able to improve Chat performance to 83.9% by performing SFT on one epoch of conversational data from OpenAssistant2 (Köpf et al., 2023).

Moreover, we believe the low quantity of high-quality, open source conversational pairwise preference data (a consequence of the prevalence of LLM-based RMs) acted as a bottleneck on our models' performance.

Reasoning Domain Specialists using DoRA achieve strong performance, suggesting that lightweight tuning methods can effectively elicit latent reasoning capabilities even in small models.

Eldan & Li (2023) suggest that it is easier for LLMs to learn grammatical structure than higher-order abilities such as creativity and reasoning. In our work, however, we surprisingly notice that our best Reasoning runs used DoRA, a low-rank finetuning method, implying there is more to reasoning ability than a direct relationship with parameters available to train. While scaling parameter count generally improves reasoning capabilities, we argue that there should be more focus on eliciting latent reasoning capabilities from rich internal representations.

Safety Domain Despite their small size, specialists generalize well to refusal tasks. As the category with the second lowest average score by leaderboard models, Safety appears to be one of the more difficult RewardBench domains.[4] This aligns with the complex nature of tasks within the Safety category–strong reasoning skills are necessary to determine whether a refusal is appropriate in a given context. With this being said, the large specialist scores in the 84% percentile of models on the leaderboard, which is similar to the Reasoning specialist's performance.

5. Conclusion and Future Work

Our findings demonstrate that bidirectional masked language models can serve as effective reward models at a fraction of the computational cost of current approaches. The TinyRM family, with models as small as 400 million parameters, achieves competitive performance with models over 175 times larger on Reasoning and Safety preference modeling tasks. Through our evaluation on RewardBench, we discovered that different domains benefit from distinct finetuning strategies—notably, reasoning tasks showed particular responsiveness to DoRA, a lightweight parameter-efficient method, suggesting that eliciting existing capabilities may be more effective than scaling parameter count.

Our results reveal both the promise and limitations of small reward models. While our specialists excel in reasoning and safety domains, they face challenges in conversational tasks, likely due to the lack of supervised fine-tuning equivalent to what larger decoder-based models receive. Nevertheless, our work establishes that bidirectional architectures combined with FLAN-style prompting, DoRA, and early layer freezing can produce substantial reward modeling capabilities in lightweight models, offering a path toward more accessible and deployable preference learning systems.

One area we hope to expand upon in future work is the reconciliation of our specialists into a single generalist model. Along the lines of previous work from Ramé et al. (2024), we employed a weight-averaging method across our specialists. However, we were unable to achieve a model that performed well across all domains.

Additionally, we are interested in the particular mechanisms through which small encoder-based models maintain such rich representations of knowledge and reasoning capabilities. The effectiveness of parameter-efficient methods like DoRA and layer freezing provide some speculative glimpses into these underlying processes, but as preliminary work, we do not provide a thorough understanding of why these lightweight approaches succeed. Moreover, analysis of failure modes may also provide information on systematic weaknesses. Lastly, we are interested in the scaling laws that govern the performance of encoder-based LMs on the reward modeling task. We make a preliminary attempt in Appendix C, but the training of more checkpoints is necessary to characterize the end behaviors of the trend line.

[4] At the time of submission.

Acknowledgements

I am grateful to the team at Answer.AI for their generous mentorship throughout this project. In particular, I thank Alexis Gallagher, Austin Huang, Benjamin Clavié, and Benjamin Warner for their invaluable insights and feedback. I am also especially thankful to Jeremy Howard for his guidance and support.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Chaffin, A., Claveau, V., and Kijak, E. Ppl-mcts: Constrained textual generation through discriminator-guided mcts decoding, 2022. URL https://arxiv.org/abs/2109.13582.

Chen, X., Li, G., Wang, Z., Jin, B., Qian, C., Wang, Y., Wang, H., Zhang, Y., Zhang, D., Zhang, T., Tong, H., and Ji, H. Rm-r1: Reward modeling as reasoning, 2025. URL https://arxiv.org/abs/2505.02387.

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences, 2023.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. Scaling instruction-finetuned language models, 2022. URL https://arxiv.org/abs/2210.11416.

Clavié, B., Cooper, N., and Warner, B. It's all in the [mask]: Simple instruction-tuning enables bert-like masked language models as generative classifiers, 2025. URL https://arxiv.org/abs/2502.03793.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Eldan, R. and Li, Y. Tinystories: How small can language models be and still speak coherent english?, 2023. URL https://arxiv.org/abs/2305.07759.

Fernandez Monsalve, I., Frank, S. L., and Vigliocco, G. Lexical surprisal as a general predictor of reading time. In Daelemans, W. (ed.), *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 398–408, Avignon, France, April 2012. Association for Computational Linguistics. URL https://aclanthology.org/E12-1041/.

Gao, L., Schulman, J., and Hilton, J. Scaling laws for reward model overoptimization, 2022. URL https://arxiv.org/abs/2210.10760.

Gao, T., Fisch, A., and Chen, D. Making pre-trained language models better few-shot learners. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3816–3830, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.295. URL https://aclanthology.org/2021.acl-long.295/.

Ghanizadeh, M. A. and Dousti, M. J. Towards data-efficient language models: A child-inspired approach to language learning, 2025. URL https://arxiv.org/abs/2503.04611.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J.,

Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Giorno, A. D., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T., and Li, Y. Textbooks are all you need, 2023. URL https://arxiv.org/abs/2306.11644.

Hou, Z., Du, P., Niu, Y., Du, Z., Zeng, A., Liu, X., Huang, M., Wang, H., Tang, J., and Dong, Y. Does rlhf scale? exploring the impacts from data, model, and method, 2024. URL https://arxiv.org/abs/2412.06000.

Howard, J. and Ruder, S. Universal language model fine-tuning for text classification, 2018. URL https://arxiv.org/abs/1801.06146.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Ji, J., Hong, D., Zhang, B., Chen, B., Dai, J., Zheng, B., Qiu, T., Li, B., and Yang, Y. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference, 2024. URL https://arxiv.org/abs/2406.15513.

Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z.-R., Stevens, K., Barhoum, A., Duc, N. M., Stanley, O., Nagyfi, R., ES, S., Suri, S., Glushkov, D., Dantuluri, A., Maguire, A., Schuhmann, C., Nguyen, H., and Mattick, A. Openassistant conversations – democratizing large language model alignment, 2023. URL https://arxiv.org/abs/2304.07327.

Lambert, N., Pyatkin, V., Morrison, J., Miranda, L., Lin, B. Y., Chandu, K., Dziri, N., Kumar, S., Zick, T., Choi, Y., Smith, N. A., and Hajishirzi, H. Rewardbench: Evaluating reward models for language modeling, 2024. URL https://arxiv.org/abs/2403.13787.

Lee, J., Tang, R., and Lin, J. What would elsa do? freezing layers during transformer fine-tuning, 2019. URL https://arxiv.org/abs/1911.03090.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019. URL https://arxiv.org/abs/1910.13461.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

Liu, C. Y., Zeng, L., Liu, J., Yan, R., He, J., Wang, C., Yan, S., Liu, Y., and Zhou, Y. Skywork-reward: Bag of tricks for reward modeling in llms, 2024a. URL https://arxiv.org/abs/2410.18451.

Liu, S.-Y., Wang, C.-Y., Yin, H., Molchanov, P., Wang, Y.-C. F., Cheng, K.-T., and Chen, M.-H. Dora: Weight-decomposed low-rank adaptation, 2024b. URL https://arxiv.org/abs/2402.09353.

Lu, K., Yuan, H., Lin, R., Lin, J., Yuan, Z., Zhou, C., and Zhou, J. Routing to the expert: Efficient reward-guided ensemble of large language models, 2023. URL https://arxiv.org/abs/2311.08692.

Luo, L., Liu, Y., Liu, R., Phatale, S., Guo, M., Lara, H., Li, Y., Shu, L., Zhu, Y., Meng, L., Sun, J., and Rastogi, A. Improve mathematical reasoning in language models by automated process supervision, 2024. URL https://arxiv.org/abs/2406.06592.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., and Schulman, J. Webgpt: Browser-assisted question-answering with human feedback, 2022. URL https://arxiv.org/abs/2112.09332.

Nogueira, R. and Cho, K. Passage re-ranking with bert, 2020. URL https://arxiv.org/abs/1901.04085.

Nvidia, :, Adler, B., Agarwal, N., Aithal, A., Anh, D. H., Bhattacharya, P., Brundyn, A., Casper, J., Catanzaro, B., Clay, S., Cohen, J., Das, S., Dattagupta, A., Delalleau, O., Derczynski, L., Dong, Y., Egert, D., Evans, E., Ficek, A., Fridman, D., Ghosh, S., Ginsburg, B., Gitman, I., Grzegorzek, T., Hero, R., Huang, J., Jawa, V., Jennings, J., Jhunjhunwala, A., Kamalu, J., Khan, S., Kuchaiev, O., LeGresley, P., Li, H., Liu, J., Liu, Z., Long, E., Mahabaleshwarkar, A. S., Majumdar, S., Maki, J., Martinez, M., de Melo, M. R., Moshkov, I., Narayanan, D., Narenthiran, S., Navarro, J., Nguyen, P., Nitski, O., Noroozi, V., Nutheti, G., Parisien, C., Parmar, J., Patwary, M., Pawelec, K., Ping, W., Prabhumoye, S., Roy, R., Saar, T., Sabavat, V. R. N., Satheesh, S., Scowcroft, J. P., Sewall, J., Shamis, P., Shen, G., Shoeybi, M., Sizer, D., Smelyanskiy, M., Soares, F., Sreedhar, M. N., Su, D., Subramanian, S., Sun, S., Toshniwal, S., Wang, H., Wang, Z., You, J., Zeng, J., Zhang, J., Zhang, J., Zhang, V., Zhang, Y., and Zhu, C. Nemotron-4 340b technical report, 2024. URL https://arxiv.org/abs/2406.11704.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L.,

Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Pan, S., Lialin, V., Muckatira, S., and Rumshisky, A. Let's reinforce step by step, 2023. URL https://arxiv.org/abs/2311.05821.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL https://arxiv.org/abs/1910.10683.

Ramé, A., Vieillard, N., Hussenot, L., Dadashi, R., Cideron, G., Bachem, O., and Ferret, J. Warm: On the benefits of weight averaged reward models, 2024. URL https://arxiv.org/abs/2401.12187.

Samuel, D. Berts are generative in-context learners, 2024. URL https://arxiv.org/abs/2406.04823.

Schick, T. and Schütze, H. Exploiting cloze questions for few shot text classification and natural language inference, 2021a. URL https://arxiv.org/abs/2001.07676.

Schick, T. and Schütze, H. It's not just size that matters: Small language models are also few-shot learners, 2021b. URL https://arxiv.org/abs/2009.07118.

Shen, L., Chen, S., Song, L., Jin, L., Peng, B., Mi, H., Khashabi, D., and Yu, D. The trickle-down impact of reward (in-)consistency on rlhf, 2023. URL https://arxiv.org/abs/2309.16155.

Skean, O., Arefin, M. R., Zhao, D., Patel, N., Naghiyev, J., LeCun, Y., and Shwartz-Ziv, R. Layer by layer: Uncovering hidden representations in language models, 2025. URL https://arxiv.org/abs/2502.02013.

Song, Z., Cai, T., Lee, J. D., and Su, W. J. Reward collapse in aligning large language models, 2023. URL https://arxiv.org/abs/2305.17608.

Steuer, J., Mosbach, M., and Klakow, D. Large gpt-like models are bad babies: A closer look at the relationship between linguistic competence and psycholinguistic measures, 2023. URL https://arxiv.org/abs/2311.04547.

Sun, C., Qiu, X., Xu, Y., and Huang, X. How to fine-tune bert for text classification?, 2020. URL https://arxiv.org/abs/1905.05583.

Wang, Z., Dong, Y., Delalleau, O., Zeng, J., Shen, G., Egert, D., Zhang, J. J., Sreedhar, M. N., and Kuchaiev, O. Helpsteer2: Open-source dataset for training top-performing reward models, 2024. URL https://arxiv.org/abs/2406.08673.

Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., Cooper, N., Adams, G., Howard, J., and Poli, I. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024. URL https://arxiv.org/abs/2412.13663.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners, 2022. URL https://arxiv.org/abs/2109.01652.

Winata, G. I., Anugraha, D., Susanto, L., Kuwanto, G., and Wijaya, D. T. Metametrics: Calibrating metrics for generation tasks using human preferences, 2025. URL https://arxiv.org/abs/2410.02381.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks?, 2014. URL https://arxiv.org/abs/1411.1792.

A. Training Details

All experiments were trained using a Decoupled AdamW optimizer with a weight decay of 1.0e-5. All runs were trained for one epoch, a batch size of 256, and a linear decay. We performed hyperparameter sweeps using a Bayesian search strategy across learning rate, DoRA rank, number of frozen layers, and a pool of instruction prefixes. The configurations that yielded the results shown in the main results table are presented below.

Table 2. Hyperparameter settings for each domain-specific reward model.

Model Size	Domain	Learning Rate	Layers Frozen	Instruction Prefix
Large				
	Chat	7.865e-5	26	"Select the best response."
	Reasoning	2.327e-5	12	"Which response is more correct?"
	Safety	5.845e-5	7	"Which response is safer?"
	All at Once	9.388e-5	5	Optimal domain-specific prompts
Base				
	Chat	7.642e-5	2	"Which response is the most helpful, relevant, and correct?"
	Reasoning	8.354e-5	17	"Select the best response."
	Safety	1.478e-5	12	"Which response is safer?"

Note: DoRA was applied only to the Reasoning specialists with rank dimension 128; all other models used standard finetuning without DoRA.

B. Training Datasets

The following datasets were used to train our specialist RMs. All of these datasets were used for the large AAO model. We used the decontaminated version of Skywork.[5]

Table 3. Datasets used for training in each domain.

Domain	Training Datasets
Chat	Skywork Chat (7.22k pairs) (Liu et al., 2024a), WebGPT Comparisons (14.3k) (Nakano et al., 2022), HH-RLHF (161k) (Bai et al., 2022)
Reasoning	Skywork Reasoning (54.6k) (Liu et al., 2024a)
Safety	Skywork Safety (15.2k) (Liu et al., 2024a) , PKU-SafeRLHF (26.9k) (Ji et al., 2024)

C. Inference Compute vs. Accuracy Tradeoff

A discriminator that is used at inference time must carry its weight in terms of the tradeoff between computation and accuracy. To this end, we use GFLOPs per token and RewardBench accuracy as metrics to compare the TinyRM specialist suite with leaderboard models. Figure 3 shows that using TinyRMs can cut inference costs by two orders of magnitude without suffering a proportionate loss in accuracy.[6]

[5]https://gist.github.com/natolambert/1aed306000c13e0e8c5bc17c1a5dd300

[6]We calculate the number of FLOPs using the approximation $\text{FLOPs} = 2\text{N}_{\text{params}} + 6\text{N}_{\text{params}}\frac{L}{H}$ where L is the number of layers and H is the hidden size. DoRA does not add significant inference-time overhead as adapters can be merged into the pretrained weights.

Figure 3. TinyRM cuts inference costs by two orders of magnitude without suffering a proportionate loss in accuracy.