RL-MR: Reinforcement Learning based Budget-Aware Sequential Pretrained Model Ranking

Saddam Mukta[a], **Md. Ahashan Habib**[b], **Mir Tafseer Nayeem**[c], **Md Adnanul Islam**[d]

[a]Software Engineering, Lappeenranta University of Technology [b]CSE Dept., United International University
[c]Computing Science, University of Alberta [d]Action Lab, Monash University

saddam.mukta@lut.fi mhabib201261@bscse.uiu.ac.bd mnayeem@ualberta.ca adnan.islam@monash.edu

Abstract

The rapid growth of pre-trained models has made selecting suitable models for downstream tasks challenging. Exhaustively fine-tuning or evaluating every candidate model is computationally expensive, while existing proxy-based metrics such as LEEP, NCE, and LogME produce one-shot static rankings that do not use feedback from earlier evaluations. We propose Reinforcement Learning–based Model Ranking (RL-MR), a budget-aware framework that reformulates pretrained model selection as a sequential decision-making problem. RL-MR learns a ranking policy that uses proxy-based transferability signals as rewards to guide the order in which candidate models are evaluated. In controlled experiments on *SST-2*, *Yelp Polarity*, *Amazon Polarity*, and *DynaSent*, RL-MR consistently improves ranking correlation over static proxy baselines, achieving up to Kendall's $\tau = 0.8695$ while maintaining Top-1 selection accuracy. These results suggest that sequential proxy feedback can provide a useful alternative to one-shot model ranking under limited evaluation budgets.

1 Introduction

Pre-trained models (PTMs), including Large Language Models (LLMs) (Zhang et al., 2023), fine-tuned on downstream tasks has become a standard paradigm for real-world deployment (Jia et al., 2022). A model zoo includes a wide range of PTMs that differ in architecture, scale, and functionality (Paszke et al., 2019). However, the rapid expansion of public model zoos creates a practical challenge: efficiently selecting suitable models for target tasks (Tan et al., 2024; Flotte et al., 2025), since exhaustively evaluating all candidate PTMs requires substantial time and computational resources.

Traditional approaches rely on manual expertise or expensive computational searches that are inaccessible to many practitioners (Pándy et al., 2022; Ding et al., 2022). A straightforward solution is to fine-tune every candidate PTM on the downstream task and select the best model (You et al., 2022), but this offers strong selection accuracy at prohibitive cost, especially for large PTMs requiring backward passes. To reduce this burden, proxy-based methods such as LEEP (Nguyen et al., 2020), NCE (Tran et al., 2019), and LogME (You et al., 2022) estimate transferability using forward-pass statistics or distributional features. These metrics provide indirect transferability signals, i.e., scores, without requiring training on the target task (Nguyen et al., 2020). Task–model relatedness can also be captured in a representation space, where ranking is performed by measuring similarity between task vectors and model vectors, as in MODEL SPIDER (Zhang et al., 2023). However, most existing approaches produce one-shot static rankings: each candidate model is scored independently, and the ranking does not use feedback from earlier evaluations. This is efficient, but it is less suited to budget-constrained selection settings, where users may evaluate only a subset of candidate models and update their choices as evidence accumulates.

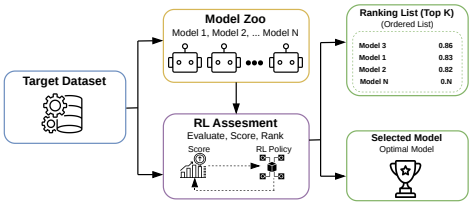


Figure 1: Overview of RL-MR for budget-aware pretrained model ranking. The policy sequentially selects models using proxy-based transferability feedback under limited evaluation budgets.

Reinforcement learning (RL) (Beck et al., 2025) provides a natural framework for studying this sequential selection setting. Static learning-to-rank (L2R) methods assign a compatibility score to each candidate independently and produce a one-shot ranking without using feedback from prior evalua-

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 540–551
November 6-10, 2026 ©2026 Association for Computational Linguistics

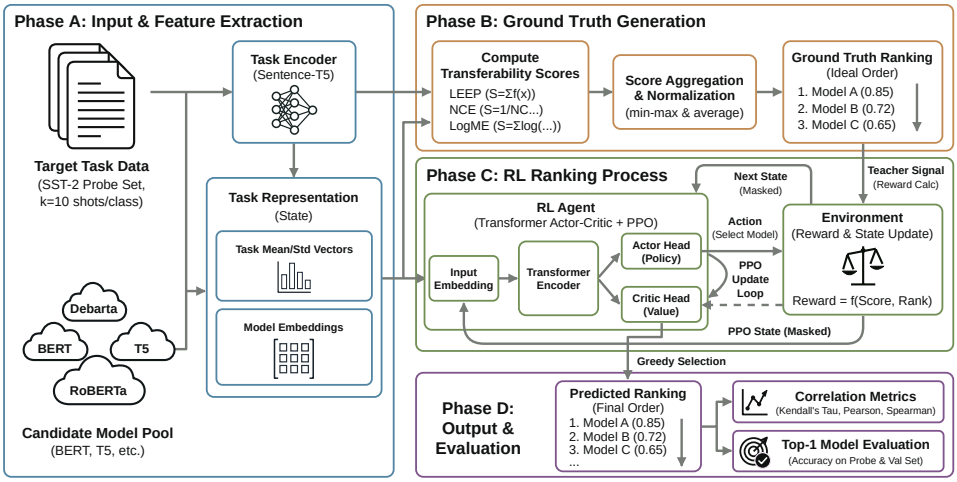


Figure 2: Phase-wise overview of RL-MR. **Phase A** encodes the target task and candidate model zoo into task–model representations; **Phase B** computes proxy-based transferability signals; **Phase C** uses these signals as feedback to update the PPO ranking policy; and **Phase D** outputs the final model ranking under a limited evaluation budget.

tions (Nguyen et al., 2020; You et al., 2022; Tran et al., 2019). We propose Reinforcement Learning–based Model Ranking (RL-MR), a budget-aware framework that reformulates pretrained model selection as a sequential decision-making problem. RL-MR does not introduce a new RL algorithm; rather, its contribution is to recast pretrained model ranking as a sequential *Markov Decision Process* (MDP), where proxy-based transferability signals provide feedback for learning a ranking policy. We instantiate this formulation using standard *Proximal Policy Optimization* (PPO) (Schulman et al., 2017). The high-level pipeline of RL-MR is shown in Figure 1. Our key contributions are as follows:

- We formulate PTM selection as a sequential RL problem rather than a one-shot static ranking task, allowing model choices to be guided by feedback from previous proxy evaluations.
- We introduce RL-MR, a PPO-based ranking policy that prioritizes promising candidate models under limited evaluation budgets without fine-tuning the PTMs.
- We show that RL-MR achieves higher Kendall-τ and Pearson-ρ correlations than static proxy-based methods under controlled sentiment-classification settings.

2 Problem Setup

Selecting suitable PTMs for a new task has become increasingly important as model zoos continue to expand (Zhang et al., 2023). Given a target classification task $T = \{(x_i, y_i)\}_{i=1}^{N}$ with C_T classes, our goal is to identify the PTM that can best transfer its knowledge to T.

Let the model zoo be represented as $\mathcal{M} = \{f_m = W_m \circ \phi_m\}_{m=1}^{M}$, where each PTM f_m consists of a feature extractor ϕ_m and a classification head W_m. Formally, we aim to estimate a set of transferability signals $\hat{t}_T = \{\hat{t}_{\phi_m \to T}\}_{m=1}^{M}$ that reflect the potential of each PTM for task T and rank models based on sequential evaluation feedback.

3 Methodology: RL-MR

We introduce RL-MR, which frames pretrained model ranking as a sequential decision-making process. The overall pipeline consists of four phases, as illustrated in Figure 2.

3.1 MDP Formulation

Given a model zoo $\mathcal{M} = \{f_m\}_{m=1}^{M}$ and a target task $T = \{(x_i, y_i)\}_{i=1}^{N}$, our goal is to rank PTMs according to their expected transferability to T. We formulate this process as an MDP. Importantly, the agent does not obtain all transferability information upfront. Evaluating a model consumes computational resources and provides only partial feedback, which influences subsequent decisions. This converts budget-constrained PTM selection into a sequential decision problem rather than a one-time ranking exercise. At each step t, the agent observes a state s_t that summarizes the task and previously evaluated models, selects an action a_t corresponding to the next model choice, and follows a policy $\pi_\theta(a_t \mid s_t)$. After evaluating the selected model, the agent receives a reward r_t reflecting its transferability, and the state is updated accordingly.

The objective of the agent is to maximize the expected cumulative return $J(\pi) = \mathbb{E}_\pi \left[\sum_{t=1}^{T} \gamma^t r_t\right]$, where γ is a discount factor that balances short-term versus long-term ranking quality. We adopt PPO (Zhuang et al., 2023) because it provides stable policy-gradient updates for sequential decision problems with continuous reward feedback. In contrast, TRPO (Kuba et al., 2021) is more expensive to optimize, while DPO (Rafailov et al., 2023) is designed for static preference alignment rather than sequential exploration.

3.2 Representation Learning

Each PTM f_m is associated with a learnable embedding $\theta_m \in \mathbb{R}^d$ that captures its semantic capability, with models trained on similar domains expected to lie close in the embedding space. The target task T is encoded using a frozen feature extractor ψ followed by class-wise centroid aggregation:

$$\mu(T)_c = \frac{1}{|I(y_i = c)|} \sum_{(x_i, y_i) \in T} \psi(x_i)\, \mathbf{1}(y_i = c),$$

producing a task representation $\mu(T) \in \mathbb{R}^{d \times C_T}$. A transformer-based module then maps concatenated representations $[\theta_m, \mu(T)]$ into a scalar score, $\hat{t}^{\phi}_{m \to T} = \mathrm{FC}(\mathrm{Transformer}([\theta_m, \mu(T)])_0)$, which estimates model–task relatedness and guides RL-based exploration.

3.3 Reward Function

To optimize sequential PTM ranking, RL-MR adopts a composite reward function that jointly considers transferability estimation, ranking consistency, and policy stability. The reward provides step-wise feedback, enabling the RL agent to refine subsequent model-selection decisions throughout the ranking trajectory.

Transferability Reward. The primary objective of RL-MR is to prioritize PTMs with high transferability to the target task. For each candidate model, proxy-based transferability scores are estimated using three complementary metrics: LEEP, NCE, and LogME. Since these metrics operate on different numerical scales, each proxy score s_i is independently normalized across the candidate model zoo:

$$\hat{s}_i = \frac{s_i - \min(s)}{\max(s) - \min(s) + \varepsilon}. \tag{1}$$

The normalized scores are aggregated as

$$S_{\text{agg}} = \frac{3\hat{s}_{\text{LEEP}} + \hat{s}_{\text{NCE}} + \hat{s}_{\text{LogME}}}{3}. \tag{2}$$

This aggregate score (S_{agg}) serves as the primary transferability signal provided to the RL agent during sequential model selection.

Ranking Reward. Besides proxy-based transferability estimation, RL-MR incorporates an additional ranking-oriented reward that encourages the policy to prioritize highly transferable models earlier in the ranking sequence:

$$r_{\text{rank}} = 1 - \frac{|\{j \in M_t : \mathrm{rank}(j) \le \mathrm{rank}(a_t)\}| - 1}{|M_t|}. \tag{3}$$

Here, M_t denotes the set of remaining candidate models at decision step t, and $\mathrm{rank}(\cdot)$ represents the ground-truth transferability order estimated from the aggregated proxy scores.

Composite Reward. The immediate reward combines the aggregate transferability estimate with the ranking-oriented reward:

$$r_t = \alpha r_{\text{trans}} + (1 - \alpha) r_{\text{rank}}. \tag{4}$$

Here, r_{trans} denotes the proxy-based transferability reward, r_{rank} represents the ranking-oriented reward, and α controls their relative contribution. The weighting coefficients were selected empirically to balance proxy-based transferability estimation and sequential ranking quality. The cumulative discounted return is

$$J(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} \gamma^t r_t\right]. \tag{5}$$

Ranking Loss. In addition to the RL objective, RL-MR incorporates a ranking loss to explicitly encourage agreement between the predicted ranking sequence and the proxy-derived target ranking:

$$L_{\text{rank}} = L_{\text{rank}}(\pi_\theta, \pi_{\text{target}}). \tag{6}$$

Here, π_θ denotes the predicted ranking produced by the policy network and π_{target} represents the proxy-derived ranking obtained from the aggregated transferability scores. Minimizing L_{rank} encourages the predicted ranking to preserve the relative ordering of candidate pretrained models. The subsequent Policy Optimization section is provided in Appendix A.3.

4 Experiments

We evaluate RL-MR on a model zoo of lightweight and medium-scale pretrained transformer models commonly used in natural language understanding, including encoder-only, encoder–decoder, multilingual, and compact architectures. The full model

zoo contains 25 language models with varying architectures and scales, as reported in Appendix Table 5. Unlike much prior work on PTM ranking, which has largely focused on vision PTMs (Zhang et al., 2023; You et al., 2022; Nguyen et al., 2020), our evaluation focuses on language models.

4.1 Experimental Setup

Datasets. We evaluate RL-MR on four widely used sentiment-classification benchmarks:*SST-2* (Socher et al., 2013), *Yelp Polarity* (Zhang et al., 2015), *Amazon Polarity* (Zhang et al., 2015), and *DynaSent* (Potts et al., 2021). These datasets provide complementary settings for assessing representation quality and ranking stability across different domains and dataset scales.

Training details. For task representation, we encode probe samples using Sentence-T5, given its strong ability to capture semantic sentence representations in few-shot settings. Each probe set contains 10 labeled examples per class, providing a practical trade-off between representativeness and efficiency. The examples are randomly drawn from the training set following prior few-shot protocols (Zhang et al., 2023). During ranking, only proxy-based forward passes on the probe set are performed; no fine-tuning or backward passes on the PTMs are used. To improve stability, we average results over 30 independent trials with different probe-set samples. Appendix Table 4 provides the full hyperparameter configuration.

Policy Training Protocol. RL-MR trains a separate PPO ranking policy for each target task rather than learning a single policy across multiple tasks. For each dataset, the agent interacts with a task-specific ranking environment constructed from the corresponding probe set and candidate model zoo. During inference, the trained policy sequentially ranks candidate pretrained models for that task without further policy updates, allowing the policy to adapt to the characteristics of each downstream task while maintaining a consistent sequential ranking framework across all experimental datasets.

Baselines. We compare RL-MR against two categories of model selection methods: i) *proxy-based methods*, which estimate transferability by relating PTM features to downstream labels, including `LEEP`, `NCE`, and `LogME`; and ii) *representation-based similarity methods*, which assess task–model compatibility in feature space, including PACTran (Ding et al., 2022), NCTI (Wang et al., 2023), MODEL SPIDER (Zhang et al., 2023), and DISCO (Zhang et al., 2025).

Metric	SST-2	Yelp	Amazon	DynaSent
NCE	-0.0815	0.0332	-0.0578	0.0423
LEEP	0.0670	-0.0420	0.1011	0.0332
LogME	0.1103	0.0718	-0.0110	-0.0160
Top-1 (FS)	0.5122 (2)	0.5018 (2)	0.5001 (1)	0.5121 (2)
Top-1 (Val.)	0.5378 (5)	0.5095 (3)	0.5005 (6)	0.5160 (3)
Kendall-τ	**0.8695 (8)**	**0.8212 (3)**	**0.8206 (3)**	**0.8623 (1)**

Table 1: Model-selection performance across four sentiment-classification benchmarks. Results for RL-MR are reported as mean (standard deviation $\times 10^{-4}$) over 30 independent trials. For example, (8) denotes a standard deviation of 0.0008.

Evaluation protocol. We report both standard evaluation on the training split and few-shot evaluation on the probe set. The key metrics are: i) *Top-1 Accuracy* measures whether the highest-ranked pretrained model selected by a ranking method matches the best-performing model identified through downstream evaluation under the same evaluation protocol. Specifically, for each trial, a prediction is considered correct if the model ranked first by the method is identical to the model achieving the highest downstream performance. We report Top-1 Accuracy under both the few-shot probe setting and the validation setting to evaluate the practical effectiveness of model selection under limited evaluation budgets (Feng et al., 2025); ii) *Ranking correlation*, measured using Kendall-τ between predicted rankings and the ground-truth ordering induced by downstream task performance under the same evaluation protocol (Zhang et al., 2023). Kendall-τ is widely used in PTM ranking studies to assess ranking-order agreement (Zhang et al., 2025; Ding et al., 2022).

4.2 Results & Analysis

Sequential feedback stabilizes ranking. Table 1 compares model-selection performance across the four sentiment benchmarks. Static proxy metrics such as NCE, LEEP, and LogME produce weak or unstable rank correlations across datasets. Their scores cluster across candidate models, so small numerical fluctuations can invert pairwise orderings and produce near-zero or negative Kendall-τ values. In contrast, RL-MR identifies stronger candidate models under both few-shot and validation settings, achieving high Kendall-τ scores across all datasets. These results suggest that sequential proxy feedback provides a more reliable signal for

Baselines	SST-2	Yelp	Amazon	DynaSent
DISCO (Zhang et al., 2025)	0.6124	0.6013	0.5879	0.5958
PACTran (Ding et al., 2022)	0.6587	0.6415	0.6721	0.6112
NCTI (Wang et al., 2023)	0.7032	0.6896	0.6948	0.6233
MODEL SPIDER (Zhang et al., 2023)	0.7425	0.7189	0.7314	0.6829
RL-MR (Ours)	**0.8695 (8)**	**0.8212 (3)**	**0.8206 (3)**	**0.8623 (1)**

Table 2: Kendall-τ comparison with representation-based baselines. Results are reported as mean (standard deviation $\times 10^{-4}$) over 30 independent trials.

pretrained model ranking, and RL-MR exploits cumulative sequential feedback to continuously refine subsequent model-selection decisions throughout the ranking trajectory rather than relying solely on one-shot proxy scores.

RL-MR improves over stronger ranking baselines. Table 2 compares RL-MR with representation-based model ranking methods using Kendall-τ correlation. RL-MR outperforms DISCO, PACTran, NCTI, and MODEL SPIDER across all four datasets. This improvement indicates that modeling model selection as a sequential decision process provides benefits beyond static task–model similarity estimation, as RL-MR uses feedback from earlier proxy evaluations to guide subsequent choices, yielding more accurate rankings under the controlled evaluation setting.

Beyond sentiment classification, we additionally evaluate RL-MR on three representative question answering benchmarks covering extractive, open-domain, and multi-hop QA. RL-MR maintains strong ranking agreement across these tasks, suggesting that the proposed sequential ranking framework generalizes beyond sentiment classification. Detailed numerical results are reported in Appendix F.

Early-budget gains are strongest. To analyze ranking efficiency under constrained computational resources, we evaluate model ranking performance as a function of the evaluation budget (Chen et al., 2023). The budget corresponds to the number of candidate models evaluated during ranking. Figure 3 shows Kendall-τ correlation as the evaluation budget increases. RL-MR achieves higher ranking correlation than static proxy metrics such as NCE, LEEP, and LogME, especially under small evaluation budgets. This indicates that the learned sequential policy prioritizes promising candidate models earlier in the ranking process. As the budget increases, the performance gap narrows, but RL-MR maintains strong ranking quality. Appendix E.1 further analyzes cumulative forward-pass cost and shows a stable growth pattern for the RL policy.

Method	Forward Eval.	PPO	Time (s)
LEEP	25	No	101.17
LogME	25	No	101.17
RL-MR	25	Yes	**823.07**

Table 3: Wall-clock runtime comparison under the same evaluation setting for ranking 25 candidate pretrained models.

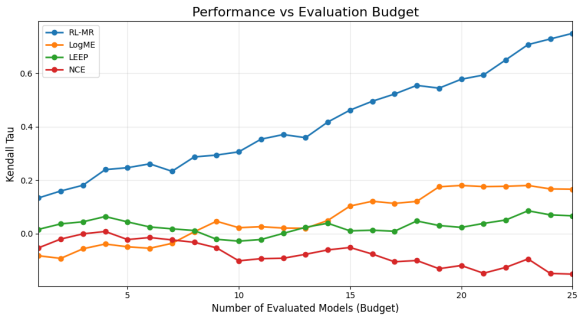


Figure 3: Performance–budget trade-off for pretrained model ranking. RL-MR achieves stronger Kendall-τ agreement under small evaluation budgets.

Computational Cost Analysis. Table 3 compares the wall-clock runtime of RL-MR with representative proxy-based ranking methods under the same evaluation setting. LEEP and LogME require 58.91 s, whereas RL-MR requires 83.55 s because of the additional PPO policy optimization stage. This additional computation is incurred only during policy learning and does not involve repeated downstream fine-tuning of candidate pretrained models. Moreover, the implementation employs batched episode processing and a unified actor–critic forward pass to reduce PPO training overhead. Although RL-MR introduces a moderate runtime increase over one-shot proxy-ranking methods, it consistently achieves substantially stronger ranking agreement across all evaluated datasets. In this work, the evaluation budget primarily refers to the expensive downstream model evaluation and adaptation budget rather than the runtime of proxy-score computation itself.

5 Conclusion

We introduced RL-MR, a reinforcement-learning framework for sequential pretrained model ranking. By combining task–model representations with policy-gradient optimization, RL-MR achieves high ranking correlation and strong selection performance under few-shot and resource-constrained settings. These results show that sequential proxy feedback can improve model selection beyond fixed one-shot proxy metrics.

Limitations

RL-MR demonstrates strong ranking agreement across sentiment classification and question answering tasks, but several limitations remain. First, RL-MR relies on proxy-based transferability scores as reward signals, which are only approximations of true downstream performance. Therefore, our results should be interpreted as evidence that sequential proxy feedback can improve model ranking, rather than as evidence that proxy signals can replace downstream validation in all settings. We also do not evaluate dynamic model pools, where candidate models may enter or leave over time, or distribution shifts across broader task families. Finally, RL-MR avoids fine-tuning candidate models, it introduces additional overhead for training the ranking policy; this trade-off should be examined more carefully at larger scale. The current study reports wall-clock runtime under a fixed evaluation setting; broader runtime analyses across larger model zoos and different hardware configurations remain important future work. Future work should evaluate RL-MR on larger candidate pools, richer cost models, downstream validation feedback, uncertainty-aware rewards, multi-objective signals, and evolving task conditions to better assess scalability and robustness.

Ethical Considerations

This work uses publicly available benchmark datasets and pretrained models, and does not collect new human-subject data or personal information. The main ethical concern is that automated model-ranking methods may influence downstream model deployment. Because proxy-based rewards can be incomplete, RL-MR should not be used as a substitute for task-specific validation, safety assessment, or fairness evaluation. We present the method as a controlled study of sequential proxy-based ranking, and future deployments should evaluate selected models under the intended use context.

References

Jacob Beck, Risto Vuorio, Evan Zheran Liu, Zheng Xiong, Luisa Zintgraf, Chelsea Finn, Shimon Whiteson, and 1 others. 2025. A tutorial on meta-reinforcement learning. *Foundations and Trends® in Machine Learning*, 18(2-3):224–384.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*.

Nan Ding, Xi Chen, Tomer Levinboim, Soravit Changpinyo, and Radu Soricut. 2022. Pactran: Pac-bayesian metrics for estimating the transferability of pretrained models to classification tasks. In *European Conference on Computer Vision*, pages 252–268. Springer.

Jie Feng, Jun Zhang, Tianhui Liu, Xin Zhang, Tianjian Ouyang, Junbo Yan, Yuwei Du, Siqi Guo, and Yong Li. 2025. Citybench: Evaluating the capabilities of large language models for urban tasks. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5413–5424.

Thomas J Flotte, Stephanie A Derauf, Rachel K Byrd, Trynda N Kroneman, Debra A Bell, Lucas Stetzik, Seung-Yi Lee, Alireza Samiei, Steven N Hart, Joaquin J Garcia, and 1 others. 2025. Democratizing artificial intelligence in anatomic pathology. *Archives of Pathology & Laboratory Medicine*, 149(1):55–59.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European conference on computer vision*, pages 709–727. Springer.

Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. 2021. Trust region policy optimisation in multi-agent reinforcement learning. *arXiv preprint arXiv:2109.11251*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. 2020. Leep: A new measure to evaluate transferability of learned representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 7294–7305. PMLR.

Michal Pándy, Andrea Agostinelli, Jasper Uijlings, Vittorio Ferrari, and Thomas Mensink. 2022. Transferability estimation using bhattacharyya class separability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9172–9182.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2021. Dynasent: A dynamic benchmark for sentiment analysis. In *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)*, pages 2388–2404.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shahil Shaik, Jonathon M Smereka, and Yue Wang. 2025. Generalized advantage estimation for distributional policy gradients. In *2025 American Control Conference (ACC)*, pages 3073–3078. IEEE.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Xin Tan, Taichuan Li, Ruohe Chen, Fang Liu, and Li Zhang. 2024. Challenges of using pre-trained models: the practitioners' perspective. In *2024 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 67–78. IEEE.

Anh Tuan Tran, Cuong V. Nguyen, and Tal Hassner. 2019. Transferability and hardness of supervised classification tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1395–1405.

Zijian Wang, Yadan Luo, Liang Zheng, Zi Huang, and Mahsa Baktashmotlagh. 2023. How far pre-trained models are from neural collapse on the target dataset informs their transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5549–5558.

Tianhao Wu, Banghua Zhu, Ruoyu Zhang, Zhaojin Wen, Kannan Ramchandran, and Jiantao Jiao. 2024. Pairwise proximal policy optimization: Language model alignment with comparative rl. In *First Conference on Language Modeling*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Kaixuan You, Yuning Liu, Zhiying Zhang, Jing Wang, Michael I Jordan, and Mingsheng Long. 2022. Ranking and tuning pre-trained models: A new paradigm for exploiting model hubs. *Journal of Machine Learning Research*, 23(209):1–47.

Tengxue Zhang, Yang Shu, Xinyang Chen, Yifei Long, Chenjuan Guo, and Bin Yang. 2025. Assessing pretrained models for transfer learning through distribution of spectral components. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 22560–22568.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Yikang Zhang, Tianjun Huang, Yuxiao Ding, Decheng Zhan, and Han-Jia Ye. 2023. Model spider: Learning to rank pre-trained models efficiently. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 13692–13719.

Mingsheng Zheng, Junwei Zhang, Changshuai Zhan, Xinyu Ren, and Shuai Lü. 2025. Proximal policy optimization with reward-based prioritization. *Expert Systems with Applications*, page 127659.

Zifeng Zhuang, Kun Lei, Jinxin Liu, Donglin Wang, and Yilang Guo. 2023. Behavior proximal policy optimization. *arXiv preprint arXiv:2302.11312*.

Supplementary Material: Appendices

A Optimization and Policy Gradient Analysis

A.1 Illustrative Example of Sequential Ranking

Assume a model zoo comprising four candidate models $\{A, B, C, D\}$ under a constrained evaluation budget. Static proxy-based methods such as LEEP calculate transferability scores independently and generate a fixed ranking as follows: $B > A > C > D$. On the contrary, RL-MR formulates the model ranking problem as a sequential decision process. At each step t, the approach selects a candidate model, estimates a transferability reward, and accordingly revises the ranking policy. For instance, selecting Model B and obtaining a high reward increases the likelihood of prioritizing semantically relevant candidates such as Model A in subsequent steps, whereas low rewards from Model D discourage similar low-potential selections. Using continuous feedback signals, RL-MR incrementally adjusts the ranking policy under constrained evaluation budgets instead of relying on a fixed single-pass scoring approach, thereby illustrating the advantage of sequential proxy-feedback-driven ranking over conventional static ranking approaches, as shown in Figure 4.

A.2 Relationship to Greedy Proxy Ranking.

Proxy-based methods such as LEEP, NCE, and LogME naturally produce deterministic one-shot rankings by sorting candidate pretrained models according to their transferability scores. Therefore, these methods serve as greedy ranking baselines without sequential feedback. In contrast, RL-MR formulates model ranking as a sequential decision-making problem, allowing the ranking policy to refine subsequent model-selection decisions based on cumulative reward feedback under a limited evaluation budget.

A.3 PPO Optimization Objective

We train the RL-MR agent using Proximal Policy Optimization (PPO) (Wu et al., 2024; Zheng et al., 2025), which constrains the probability ratio between new and old policies to support stable updates. PPO serves as the policy-gradient optimizer for learning the ranking policy. At each decision step, the policy predicts the next candidate model to be selected based on the current environment state, receives an immediate reward, and stores the transition in a rollout buffer. After completing each ranking trajectory, the collected experiences are used to update the actor and critic networks through multiple PPO optimization epochs. Given the policy $\pi_\theta(a_t \mid s_t)$, the agent maximizes the expected discounted return

$$J(\pi_\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=1}^{T} \gamma^t r_t\right],$$

where $\gamma \in [0, 1)$ is the discount factor. The advantage is estimated as

$$\hat{A}_t = R_t - V_\phi(s_t), \quad R_t = \sum_{k=0}^{T-t} \gamma^k r_{t+k},$$

where $V_\phi(s_t)$ denotes the value function. The critic is trained by minimizing the mean-squared error between predicted and target returns:

$$\mathcal{L}_{\text{critic}}(\phi) = \mathbb{E}_t\left[(V_\phi(s_t) - R_t)^2\right].$$

The PPO policy objective is defined as

$$\begin{aligned} L_{\text{PPO}} = \mathbb{E}\Big[& \min\left(r_t(\theta)\hat{A}_t,\right. \\ & \left.\text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\,\hat{A}_t\right)\Big], \end{aligned} \tag{7}$$

where $r_t(\theta)$ denotes the probability ratio between the updated and previous policies, $\hat{A}_t$ represents the generalized advantage estimate (GAE) (Shaik et al., 2025), and ϵ is the clipping coefficient that stabilizes policy updates by preventing excessively large optimization steps.

The value loss is given by

$$L_{\text{value}} = \text{MSE}(V(s_t), R_t). \tag{8}$$

We also include an entropy regularizer to encourage exploration and prevent premature convergence:

$$\mathcal{L}_{\text{entropy}} = -\beta\,\mathbb{E}_t\left[H(\pi_\theta(a_t \mid s_t))\right],$$

where β controls the strength of exploration.

The final optimization objective jointly combines the PPO policy objective, critic regression loss, entropy regularization, and the proposed ranking loss. While the PPO objective provides sequential reward-based feedback, the ranking loss encourages agreement between the predicted ranking and the proxy-derived target ranking, thereby improving ranking consistency.

The reward signal driving this optimization is defined by the composite reward introduced in

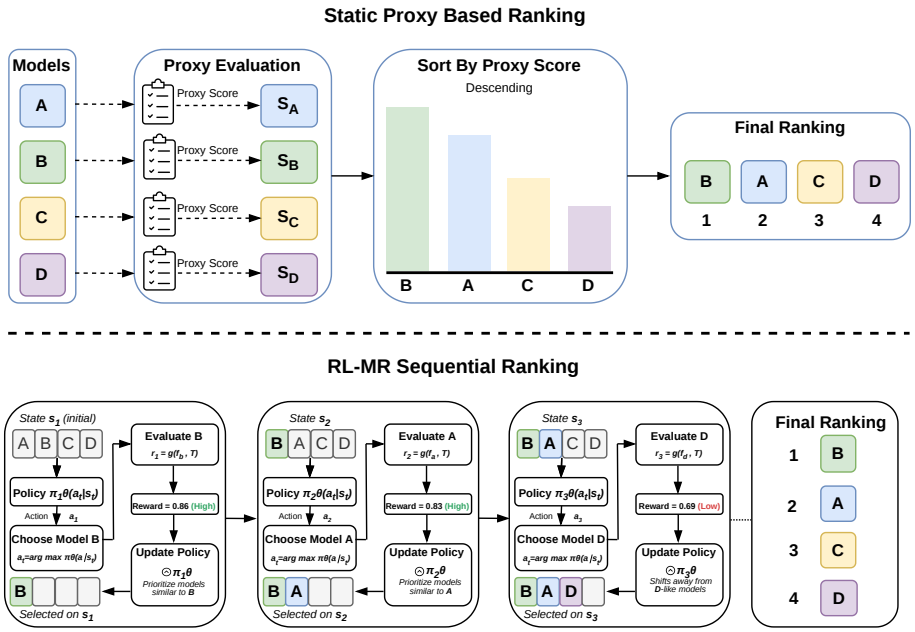


Figure 4: Illustration of sequential ranking in RL-MR. Unlike static proxy-based ranking methods that produce a fixed ordering, RL-MR iteratively updates model selection decisions through reward-driven feedback under constrained evaluation budgets.

Section 3.3, which combines proxy-based transferability estimation with ranking-oriented feedback. Thus, PPO optimizes the ranking policy using the step-wise rewards generated during the sequential model-selection process.

B Algorithm

Algorithm 1 summarizes the complete PPO optimization procedure. At each interaction step, the agent selects one candidate model, receives the corresponding reward, updates the environment state, and stores the transition for subsequent PPO optimization.

C Hyperparameter Settings

Table 4 summarizes the PPO optimization and reward-related hyperparameters used in all experiments. These settings are fixed across datasets.

D Model Zoo Details

Table 5 lists the 25 language models in our model zoo. The set includes parameter-efficient encoders, encoder–decoder models, multilingual models, and compact architectures, providing variation in size, architecture, and pretraining objective.

Algorithm 1 Reinforcement Learning–based Model Ranking (RL-MR) with PPO

Require: Model zoo $\mathcal{M}$, task distribution $\mathcal{T}$, reward function $r(\cdot)$, policy π_θ, value network V_ϕ, hyperparameters $(\gamma, \lambda_v, \beta, \epsilon, \alpha)$

for each training episode **do**
  Sample task $\mathcal{T}_i \sim P(\mathcal{T})$
  Initialize state $s_0 \leftarrow \mu(\mathcal{T}_i)$
  Initialize trajectory buffer $\mathcal{B} \leftarrow \varnothing$
  for $t = 1$ to T **do**
    Select action $a_t \sim \pi_\theta(a_t \mid s_t)$
    Compute reward $r_t = r(a_t, \mathcal{T}_i)$
    Observe next state s_{t+1}
    Store (s_t, a_t, r_t, s_{t+1}) in $\mathcal{B}$
  end for
  Compute returns $R_t = \sum_{k=0}^{T-t} \gamma^k r_{t+k}$ and advantages $\hat{A}_t = R_t - V_\phi(s_t)$
  for each PPO update iteration **do**
    $r_t(\theta) = \dfrac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}$
    Compute policy loss:

$$\mathcal{L}_{\text{PPO}} = \mathbb{E}_t\Big[\min\big(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\big)\Big]$$

    Critic loss: $\mathcal{L}_{\text{critic}} = \mathbb{E}_t[(V_\phi(s_t) - R_t)^2]$
    Entropy bonus: $\mathcal{L}_{\text{entropy}} = -\mathbb{E}_t[H(\pi_\theta(\cdot \mid s_t))]$
    Ranking loss: $\mathcal{L}_{\text{rank}}$
    Total loss:

$$\mathcal{L}_{\text{total}} = -\mathcal{L}_{\text{PPO}} + \lambda_v \mathcal{L}_{\text{critic}} - \beta \mathcal{L}_{\text{entropy}} + \alpha \mathcal{L}_{\text{rank}}$$

    Update parameters $\theta, \phi \leftarrow \text{Adam}(\nabla_{\theta,\phi} \mathcal{L}_{\text{total}})$
  end for
end for

Hyperparameter	Value
Discount factor γ	0.99
GAE parameter λ	0.95
PPO clipping ϵ	0.2
PPO epochs	8
Episodes per epoch	64
Total training epochs	30
Minibatch size	128
Learning rate	1×10^{-5}
Entropy coefficient β	0.01
Transfer reward weight	0.7
Consistency reward weight	0.3
Probe shots per class	10

Table 4: Fixed hyperparameters used for PPO optimization and reward weighting across all datasets.

Category	Models
Small Baselines	`bert-tiny`, `bert-mini`, `distilbert-base-uncased`, `distilroberta-base`
Encoder-only	`bert-base-uncased`, `bert-large-uncased`, `roberta-base`, `roberta-large`, `deberta-v3-base`, `deberta-v3-large`, `electra-base`, `albert-base-v2`
Encoder-Decoder	`bart-base`, `bart-large`, `t5-small`, `t5-base`, `t5-large`, `flan-t5-small`, `flan-t5-base`
Multilingual	`xlm-roberta-base`, `xlm-roberta-large`
Compact Models	`deberta-v3-small`, `bert_L-4_H-512_A-8`, `squeezebert-uncased`

Table 5: Model zoo used in the experiments, grouped by architecture type. The 25 candidates cover small baselines, encoder-only models, encoder–decoder models, multilingual models, and compact architectures.

E Additional Experimental Results

E.1 Forward-pass Cost Analysis

Figure 5 reports the cumulative forward-pass evaluation cost as RL-MR explores candidate models. The cost increases progressively as additional candidates are evaluated, reflecting the sequential allocation of evaluation resources. The smooth growth pattern indicates stable evaluation behavior without excessive concentration on a small subset of candidates. RL-MR also prioritizes informative models early in the ranking process, supporting strong ranking agreement under constrained evaluation budgets.

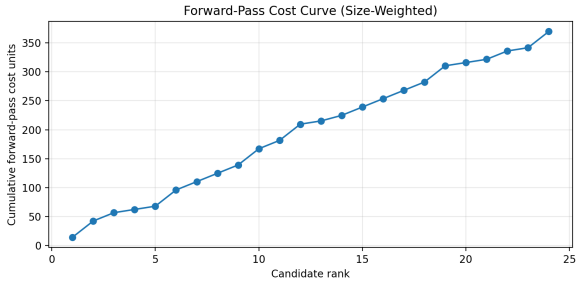


Figure 5: Cumulative forward-pass cost during RL-MR ranking. Cost increases smoothly with candidate position, indicating stable sequential allocation of proxy evaluations.

Dataset	Pearson	Spearman's Footrule
SST-2	0.9539	0.8311
Yelp Polarity	0.9045	0.8450
Amazon Polarity	0.9122	0.8356
DynaSent	0.9634	0.8367

Table 6: Extended ranking agreement for RL-MR across datasets. Pearson-ρ correlation measures linear agreement with the ground-truth ranking, while Spearman's Footrule captures rank-order discrepancy.

E.2 Extended Ranking Correlation Analysis

Beyond the top-1 model selected by each baseline metric, we assess how well the complete RL-MR ranking matches the ground-truth ordering of pretrained models. Following prior ranking work (Wang et al., 2023; Ding et al., 2022), we report *Pearson correlation* and *Spearman's Footrule* to capture agreement between predicted and ground-truth rankings (Table 6). Across all four datasets, RL-MR shows strong agreement with the true transferability ordering, further supporting the benefit of sequential model selection.

E.3 Training Dynamics of Ranking Agreement

Figure 6 shows the evolution of ranking agreement during RL-MR training. We report Kendall-τ and Pearson-ρ correlation between the ranking predicted by the RL policy and the ground-truth ranking obtained from downstream validation performance. RL-MR improves rapidly during the early stages of PPO optimization and gradually converges to stable high-correlation solutions. Pearson-ρ correlation reaches approximately 0.95, indicating strong linear agreement, while Kendall-τ stabilizes near 0.86, indicating consistent pairwise ranking agreement. These results show that the sequential RL framework progressively aligns the

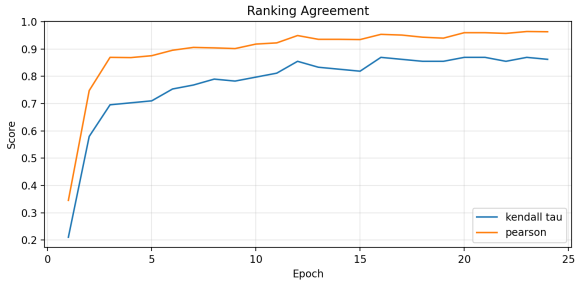


Figure 6: Training dynamics of RL-MR ranking agreement. Kendall-τ and Pearson-ρ correlation improve during PPO optimization, indicating convergence toward the ground-truth transferability ordering.

Epoch	SST-2		DynaSent	
	τ	Top-1	τ	Top-1
1	0.2101	×	0.4928	×
5	0.7101	×	0.7609	✓
10	0.7971	✓	0.7899	✓
15	0.8188	✓	0.8478	✓
20	**0.8696**	✓	**0.8623**	✓
Best	**0.8696**	✓	**0.8623**	✓

Table 7: Epoch-wise training dynamics of RL-MR on SST-2 and DynaSent across 25 candidate models. Kendall-τ improves progressively from near-random correlation at epoch 1 to peak performance at the best epoch. Top-1 accuracy (✓ = correct model selected, × = incorrect) stabilizes correctly by epoch 9 on SST-2 and epoch 3 on DynaSent. Yelp Polarity and Amazon Polarity follow similar convergence patterns as shown in Figure 6.

policy with the true transferability ordering of pretrained models.

To better understand this convergence trend in finer detail, we report both ranking correlation and model selection accuracy at each training epoch. Table 7 provides an epoch-level breakdown of training progression, tracking both ranking correlation and Top-1 accuracy, and showing that the policy improves from near-random correlation (τ=0.21) at epoch 1 to τ=0.87 at convergence on SST-2.

E.4 Reward Component Ablation

To assess the contribution of each reward component, we evaluate three configurations: (i) transfer reward only, (ii) transfer plus consistency reward, and (iii) the full composite reward including the regularization term. Table 8 reports Kendall-τ across all four datasets. Adding the consistency reward improves average Kendall-τ by +0.062 over transfer-only, which confirms that penalizing large deviations from prior ranking decisions stabilises

Configuration	SST-2	Yelp	Amazon	DynaSent
Transfer only	0.7821	0.7304	0.7198	0.7756
Transfer + consistency	0.8412	0.7965	0.7888	0.8301
Full reward (RL-MR)	**0.8695**	**0.8212**	**0.8206**	**0.8623**

Table 8: Reward ablation: Kendall-τ ranking correlation across reward configurations.

the learned policy. Including the regularization term further improves correlation by +0.029, suggesting it prevents overconfident policy updates in low-budget settings.

F Additional Evaluation of RL-MR on Question Answering Tasks

Table 9 extends the evaluation of RL-MR beyond sentiment classification to more challenging question answering (QA) settings, including *extractive*, *open-domain*, and *multi-hop reasoning* tasks. We utilized three popular datasets for QA downstream task: *SQuAD* (Rajpurkar et al., 2018), *Questions* (Kwiatkowski et al., 2019) and *HotpotQA* (Yang et al., 2018). Across all datasets, RL-MR demonstrates consistently better ranking agreement, with Kendall-τ values exceeding 0.82 and Pearson-ρ values above 0.91, which includes that the sequential ranking policy effectively captures the transferability relationships between pretrained models and downstream QA tasks. We utilized the following ten models for ranking on QA datasets: BERT-based Models (*BERT, RoBERTa, ALBERT, SpanBERT*), Encoder-Decoder Models (T5, BART, UnifiedQA), QA-Specific Models (ELECTRA, DeBERTa, XLNet). Performance on *SQuAD* remains the strongest due to its relatively structured extractive setting, while *Natural Questions* and *HotpotQA* present additional challenges related to retrieval complexity and multi-step reasoning. Although performance decreases moderately on these harder benchmarks, RL-MR continues to prioritize highly transferable models under constrained evaluation budgets. These results suggest that incorporating sequential proxy feedback enables more stable and reliable pretrained model ranking across diverse question answering scenarios.

G Broader Connections to Related Research Areas

Beyond its application to PTM ranking, the RL-MR framework may generalize to several broader research dimensions. First, RL-MR can be framed

Dataset	Task	Kendall-τ	Pearson-ρ
SQuAD	Extractive QA	0.8512	0.9345
Natural Questions	Open-domain QA	0.8234	0.9121
HotpotQA	Multi-hop QA	0.8389	0.9257

Table 9: RL-MR evaluation across extractive and reasoning-based question answering datasets. Results are reported using Kendall-τ and Pearson-ρ metrics to assess transferability-aware ranking agreement.

as a sequential policy learning problem, revealing its close alignment with meta-learning and meta-reinforcement learning paradigms (Beck et al., 2025), in which agents use feedback to improve future decision-making. While conventional meta-learning primarily focuses on parameter adaptation, RL-MR is explicitly designed for acquiring adaptive model selection strategies under constrained evaluation budgets. Second, RL-MR could potentially be applied to adaptive evaluation and resource-constrained inference systems, in which feedback signals are used for dynamic allocation of computational resources. In RL-MR, proxy-based transferability estimates guide the evaluation of sequential decisions, enabling the selection of promising models under limited evaluation budgets. Finally, RL-MR shares conceptual similarities with lifelong and continual learning settings, where the nature of model repositories and downstream task distributions evolves over time. In such scenarios, RL-MR can be extended to dynamic environments in which models are continuously added, removed, or updated.