Cross-Environment Neural Reranking for Sample-Efficient Action Selection in Text-Based Agents

Anonymous ACL submission

Abstract

Large language model agents achieve strong performance on text-based benchmarks but incur prohibitive inference costs, motivating the use of compact neural rerankers for action selection. We investigate whether a single lightweight model can perform action selection *across* multiple diverse environments—a capability that would eliminate per-environment model maintenance. Training DeBERTa-v3 (184M–434M parameters) jointly on ALFWorld, WebShop, and ScienceWorld with minority-class upsampling, we find that rebalanced two-environment joint training more than doubles single-environment ALFWorld performance (net gain $+0.696$, unified evaluation) while preserving WebShop performance ($+0.237$ vs. $+0.242$ single-environment). Three-environment training yields a mean combined net gain of $+0.611 \pm 0.020$ across 4 seeds, with per-environment results matching or exceeding specialized single-environment models. Cross-environment adaptation is highly sample-efficient: fine-tuning on only 9.2% of target-domain data recovers 93% of full-data performance, and scaling from base to large yields consistent improvements. Environment-aware LoRA adapter routing with PCGrad achieves a best-seed result of $+0.689$ but exhibits high variance (3-seed mean $+0.566 \pm 0.234$), representing a promising but unstable direction. Data rebalancing is the critical ingredient. We release our three-environment benchmark of 51,580 states and all model checkpoints.

1 Introduction

Recent advances in large language models have produced agents that achieve strong performance on individual benchmarks such as ALFWorld (Shridhar et al., 2021), WebShop (Yao et al., 2022a), and ScienceWorld (Wang et al., 2022). Yet these successes are fragile: an agent trained for household tasks cannot navigate an e-commerce website, and a model specialized for web search fails on science experiments. This brittleness stems from a fundamental tension between *specialization*—which maximizes within-domain performance—and *generalization*—which requires representations that transfer across domains.

This tension is compounded by cost. A single ReAct-style (Yao et al., 2022b) action decision on WebShop consumes approximately 2,000 tokens of LLM context, translating to roughly $13,000 in inference costs for a 1,500-episode evaluation run. LLM agents can also produce syntactically invalid or contextually implausible actions (Yao et al., 2022b; Shinn et al., 2024). Deploying such agents at scale is economically impractical.

Compact neural rerankers offer a practical alternative: they score pre-enumerated candidate actions, achieving comparable selection accuracy at orders-of-magnitude lower cost (Chen, 2024). A DeBERTa-v3-base reranker (He et al., 2023) (184M parameters) performs a forward pass in under 10ms on an A100 GPU at a fraction of a cent per thousand decisions. The open question is whether compact models can perform action selection *across environments* without the catastrophic specialization that has characterized prior work.

Prior work has focused almost exclusively on *within-environment* evaluation (Chen, 2024; Niu et al., 2024; Xiang et al., 2024). The central, underexplored question is: **Can a neural reranker trained jointly on multiple diverse environments achieve positive cross-environment transfer?** If shared representations emerge during joint training: (a) a single model should match or exceed

environment-specific models; (b) data imbalance should be correctable through simple rebalancing; and (c) fine-tuning on a new domain should require substantially fewer samples than training from scratch.

We systematically investigate these hypotheses across three text-based environments spanning fundamentally different domains (household, e-commerce, science), action spaces ($\sim$50 to $> 10^4$ actions), and observation formats. Our core contributions are:

1. **Rebalanced joint training enables positive cross-environment transfer.** Naively merging multi-environment data causes majority domains to dominate gradients. Minority-class upsampling ($6\times$ for ALFWorld) corrects this: the rebalanced joint model attains $+0.696$ net gain on ALFWorld—more than doubling the single-environment result—while matching WebShop single-environment performance ($+0.237$ vs. $+0.242$). Extending to three environments, the joint model achieves a mean combined net gain of $+0.611 \pm 0.020$ across 4 seeds, with per-environment results matching or exceeding single-environment baselines (Sections 5.1–5.5).

2. **Few-shot adaptation and model scaling.** Fine-tuning an ALFWorld-pretrained reranker on only 9.2% of WebShop training data recovers 93% of full-data performance; 20.1% reaches 97% of the ceiling, demonstrating reusable cross-environment representations. Scaling from DeBERTa-v3-base (184M) to large (434M) yields consistent but sublinear gains ($+7.6\%$ for $2.4\times$ parameters), indicating data diversity, not capacity, is the primary bottleneck (Sections 5.3, 5.4).

3. **Environment-aware adapter routing: promise and instability.** Extending the reranker with LoRA adapter routing and PCGrad gradient surgery achieves a best-seed combined net gain of $+0.689$, a 12.8% improvement over the three-environment baseline mean. However, replication reveals high variance (seed 123 collapses to $+0.296$, seed 456 achieves $+0.712$; 3-seed mean $+0.566 \pm 0.234$), indicating that learned routing is currently unstable and sensitive to initialization (Section 5.6).

4. **Benchmark and data release.** We construct a unified candidate-format dataset spanning 51,580 states (455,473 examples) with variable-size candidate sets (1 expert + dynamically sampled negatives; average 8.83 candidates per state). For ScienceWorld, we combine human-written trajectories (18,397 steps from ETO (Yuan et al., 2024)) with oracle simulator rollouts (8,461 steps), providing the largest public ScienceWorld action-selection dataset. We release all data, model checkpoints, and evaluation scripts.

2 Related Work

2.1 Text-Based Agents and Benchmarks

ALFWorld (Shridhar et al., 2021), WebShop (Yao et al., 2022a), and ScienceWorld (Wang et al., 2022) are primary testbeds for language-grounded agents, with complementary demands: spatial reasoning with constrained action spaces ($\sim$50 templates), compositional attribute matching over $> 10^4$ actions, and multi-step causal reasoning across 30 task types. LLM-based agents with chain-of-thought reasoning (Yao et al., 2022b) and self-reflection (Shinn et al., 2024) achieve strong individual results but at substantial inference cost. World-model-augmented approaches like WKM (Qiao et al., 2024) demonstrate cross-task transfer via LoRA modules on 7B+ LLMs, while neurosymbolic methods like EXPLORER (Basu et al., 2024) require hand-crafted rule templates. AgentBench (Liu et al., 2024) standardizes evaluation across eight environments but does not study cross-environment training dynamics.

2.2 Action Selection via Reranking

Learned scoring functions that rank pre-generated candidates offer a pragmatic middle ground between behavioral cloning and LLM generation. JudgeRank (Niu et al., 2024) uses LLM-based reasoning chains but incurs multi-step query latency. Prospector (Kim et al., 2024) ranks entire trajectories, solving a complementary problem to per-step scoring. Retrospex (Xiang et al., 2024) combines LLM likelihoods with offline RL critic scores but is evaluated only within single environments. Critically, *none* of these methods examines whether joint training across environments yields positive transfer—if it fails, practitioners must maintain separate models per environment, multiplying deployment cost.

2.3 Cross-Environment Generalization

Cross-environment transfer for text-based agents is underexplored. CLIN (Majumder et al., 2024) learns causal abstractions across ScienceWorld tasks but requires a frozen LLM. CoPS (Yang et al., 2024) enables cross-task experience sharing with theoretical guarantees, evaluated on ALFWorld and WebShop. WKM (Qiao et al., 2024) achieves cross-task transfer on all three environments we study, but its knowledge model is a LoRA adapter on a 7B LLM—we ask whether similar benefits can be realized with *two orders of magnitude fewer parameters*. Our work provides the first systematic study of data rebalancing, few-shot fine-tuning, and model scaling in cross-environment reranking with compact models (184M–434M parameters).

3 Method

3.1 Problem Formulation

We formulate action selection as a ranking problem. At each timestep t, the agent observes a textual state o_t and a task description g. A candidate set $\mathcal{C}_t = \{a_1, \ldots, a_K\}$ is provided, where exactly one candidate $a^* \in \mathcal{C}_t$ is the correct (expert) action and the remaining $K - 1$ are distractors sampled from other states. Candidate set sizes vary (2–20, average 8.83).

A reranker f_θ parameterized by θ maps each candidate to a scalar score:

$$s_k = f_\theta(o_t, g, a_k), \quad a_k \in \mathcal{C}_t \tag{1}$$

The agent selects $\hat{a} = \arg\max_k s_k$, and the reranker is trained to assign the highest score to the expert action a^*.

3.2 Feature Representation

We adopt a `state_action_overlap` feature mode, which constructs a structured text input by concatenating the observation, candidate action, and lexical overlap features:

$$\begin{aligned}\phi(o, a) = {} & \texttt{state: } o \\ & \oplus \texttt{\textbackslash n action: } a \\ & \oplus \texttt{\textbackslash n features: } \psi(o, a)\end{aligned} \tag{2}$$

where $\psi(o, a)$ extracts word-level overlap statistics including Jaccard coefficient, token coverage ratios, and bigram overlap between the observation text and the action string. These features provide explicit lexical alignment signals that complement the encoder's learned representations.

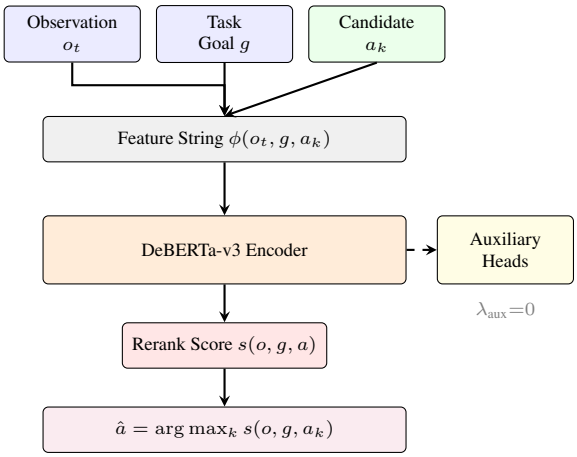

Figure 1: Architecture overview. The observation, task goal, and each candidate action are combined into a feature string with lexical overlap statistics. A DeBERTa-v3 encoder produces a representation from which the rerank head computes a scalar score. Auxiliary prediction heads (dashed) are available but disabled in our main experiments.

3.3 Model Architecture

Our architecture consists of a DeBERTa-v3 encoder (He et al., 2023) followed by a linear reranking head. Given the feature string $\phi(o, a)$, the encoder produces a contextualized representation:

$$\mathbf{h} = \text{DeBERTa-v3}(\phi(o, a)) \tag{3}$$

The rerank score is computed from the `[CLS]` token representation:

$$s(o, a) = \mathbf{w}_r^\top \mathbf{h}_{\texttt{[CLS]}} + b_r \tag{4}$$

We experiment with two encoder scales:

- **DeBERTa-v3-base**: 184M parameters (12 layers, hidden size 768, 12 attention heads)
- **DeBERTa-v3-large**: 434M parameters (24 layers, hidden size 1024, 16 attention heads)

Optionally, the model can include auxiliary prediction heads for multi-task learning (action source classification, goal-slot prediction, object and receptacle identification, etc.). In our main experiments, we disable auxiliary losses ($\lambda_{\text{aux}} = 0$) to isolate the effect of the ranking objective.

3.4 Training Objective

The model is trained with a composite loss combining pairwise ranking and pointwise classification objectives.

Pairwise ranking loss. For a positive action a^+ and a set of negative actions $\mathcal{N}$, we minimize the hinge loss:

$$\mathcal{L}_{\text{pair}} = \frac{1}{|\mathcal{N}|} \sum_{a^- \in \mathcal{N}} \ell(a^+, a^-),$$
$$\ell(a^+, a^-) = \max\big(0,\ m - s(o, a^+) + s(o, a^-)\big). \tag{5}$$

We set the margin $m = 0$; increasing the margin did not improve results in preliminary experiments.

Pointwise classification loss. As an auxiliary signal, we apply binary cross-entropy between the predicted score (passed through a sigmoid) and the ground-truth label ($y = 1$ for the expert action, $y = 0$ for negatives):

$$\mathcal{L}_{\text{point}} = -\left[y \log \sigma(s) + (1 - y) \log(1 - \sigma(s))\right] \tag{6}$$

The final loss is a weighted sum:

$$\mathcal{L} = \mathcal{L}_{\text{pair}} + \lambda_{\text{point}} \mathcal{L}_{\text{point}} \tag{7}$$

with $\lambda_{\text{point}} = 0.5$ throughout. We found this weighting to provide stable training and consistent convergence across all environments.

3.5 Evaluation Metric

Our primary metric is **net gain**: the absolute improvement in the top-1 expert action selection rate achieved by the reranker over the original candidate ordering:

$$\text{Net Gain} = \frac{1}{N} \sum_{i=1}^{N} \Big[\mathbb{1}(\hat{a}_i = a_i^*) - \mathbb{1}\Big(a_i^{(0)} = a_i^*\Big)\Big] \tag{8}$$

where $\hat{a}_i$ is the top-ranked candidate after reranking and $a_i^{(0)}$ is the first candidate in the original ordering. Since candidates are randomly shuffled, the expected original top-1 rate is $1/|\mathcal{C}_t|$ (approximately 12.5% on average), providing a well-calibrated baseline.

4 Experimental Setup

4.1 Environments and Data Construction

We construct candidate datasets from three text-based environments. For each environment, we obtain oracle trajectories (expert action sequences) and generate negative candidates per state by dynamically sampling actions from other states and task types (variable-size candidate sets, average 8.83 per state, range 2–20).

Table 1: Environment statistics. State counts reflect training data after upsampling (ALFWorld 6×). Each state is an observation-action pair paired with negative distractors; the combined dataset contains 455,473 candidate examples across 51,580 states.

Environment	Domain	States (post-upsample)	Avg. Actions per Episode
ALFWorld	Household tasks	11,808	6.5
WebShop	E-commerce search	12,914	4.4
ScienceWorld	Scientific reasoning	26,858	15.3
Total	—	**51,580**	—

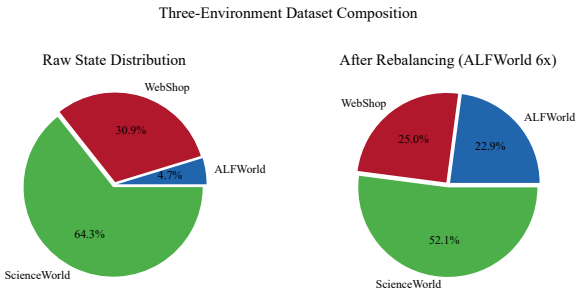

Figure 2: Dataset composition before and after rebalancing. Left: raw distribution (ALFWorld: 1,968; WebShop: 12,914; ScienceWorld: 26,858). Right: after 6× ALFWorld upsampling (ALFWorld: 11,808; WebShop: 12,914; ScienceWorld: 26,858).

ALFWorld. 100 oracle episodes (1,968 unique states before upsampling, 11,808 after 6× upsampling), negative candidates from stratified sampling (same task type, other task types, global pool).

WebShop. 1,572 human demonstration trajectories into candidate format. 12,914 states.

ScienceWorld. Two complementary sources: 18,397 steps from the ETO SFT dataset (Yuan et al., 2024) (ShareGPT-format, 1,483 episodes across 30 task types) and 8,461 oracle simulator rollouts (10 variations per task type). Combined 26,858 states—the largest public ScienceWorld action-selection dataset.

4.2 Training Protocol

All models use AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$), learning rate 2×10^{-5} to 3×10^{-4}, weight decay 0.01, linear warmup (5% steps), gradient clipping at norm 1.0. Inputs capped at 256 subword tokens, 3 epochs, 80/20 episode-level train/test split. Training on a single A100 40GB GPU. For joint training, batches sample proportionally from all environments; minority environments are upsampled ($U_{\text{ALF}} = 6$, $U_{\text{WS}} = 1$, $U_{\text{SW}} = 1$).

Table 2: Cross-environment evaluation of DeBERTa-v3-base rerankers. Net gain values (higher is better). Positive values indicate improvement over the random baseline (∼12.5% top-1 rate). All numbers from unified evaluation protocol (`evaluate_all_models.py`, episode split, seed 42).

Training Strategy	ALFWorld	WebShop	Characterization
ALFWorld only	+0.341	−0.021	Catastrophic transfer
WebShop only	+0.114	**+0.242**	Moderate transfer
Joint (equal weight)	+0.546	+0.227	Strong positive transfer both ways

5 Results

5.1 Joint Training Outperforms Single-Environment Training

We first establish baseline results by training DeBERTa-v3-base rerankers on individual environments and testing their cross-environment performance.

Several patterns emerge from Table 2. First, single-environment models transfer poorly: the ALFWorld-only reranker achieves +0.341 on its training domain but *reduces* WebShop performance slightly below random chance (−0.021). Conversely, WebShop-only training yields only 33% of the ALFWorld in-domain unified performance (+0.114 vs. +0.341). This confirms that environment-specific action distributions and observation formats are sufficiently different to preclude zero-shot transfer.

Second, joint training with equal environment weights achieves *strong positive* net gains on both domains simultaneously (+0.546 ALFWorld, +0.227 WebShop), already exceeding single-environment WebShop performance (+0.242) and substantially outperforming any single-environment model on its out-of-domain task. This indicates that shared encoder representations capture useful signals across environments. This confirms a fundamental trade-off between specialization and cross-domain robustness that joint training helps to mitigate.

5.2 Rebalanced Sampling Closes the Performance Gap

Under unified evaluation, the equal-weight joint model already outperforms ALFWorld self-training on ALFWorld data (+0.546 vs. +0.341), indicating positive cross-environment transfer from WebShop data. We hypothesize that the residual gap stems primarily from data imbalance: the 1:6.5 ratio of ALFWorld to WebShop unique states causes the minority environment to be underrepresented.

Table 3: Rebalanced joint training (6× ALFWorld upsampling) vs. baselines. All models use DeBERTa-v3-base. All numbers from unified evaluation protocol (`evaluate_all_models.py`, episode split, seed 42, `merged_candidates_rebalanced.jsonl`).

Model	ALFWorld	WebShop	Combined
Single-environment (best)	+0.341	+0.242	—
Joint (equal weight)	+0.546	+0.227	+0.390
Joint (rebalanced)	**+0.696**	**+0.237**	**+0.472**

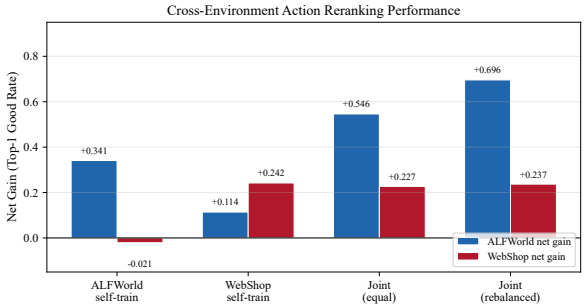

Figure 3: Net gain comparison across training strategies. The rebalanced joint model (6× ALFWorld) achieves positive gains on both environments simultaneously, more than doubling the single-environment ALFWorld ceiling and matching the WebShop ceiling.

Table 3 confirms our hypothesis. With 6× ALFWorld upsampling, the ALFWorld net gain reaches +0.696—more than *doubling* the single-environment ALFWorld performance under unified evaluation (+0.341) and improving substantially over the equal-weight joint baseline (+0.546). This comes at essentially no cost to WebShop: +0.237 vs. +0.227 (equal-weight joint), achieving 98% of the WebShop single-environment performance (+0.242). The aggregate net gain of +0.472 is 21% higher than equal-weight joint training (+0.390).

5.3 Few-Shot Fine-Tuning Is Highly Sample-Efficient

We take an ALFWorld-pretrained DeBERTa-v3-base reranker and fine-tune on increasing subsets of WebShop data (full training set: 10,456 states). Zero-shot transfer fails (−0.016), but only 9.2% of data (965 states) suffices for 93% of full-data performance (+0.201 vs. +0.217); 20.1% reaches 97% (+0.211). The saturation curve indicates that the encoder learns reusable cross-environment features, needing only modest target-domain adaptation (see Appendix A for full table and figure).

5.4 Model Scaling Yields Consistent but Sublinear Gains

DeBERTa-v3-large (434M, batch size 8) achieves combined net gain $+0.508$ under unified evaluation on the rebalanced two-environment dataset, a $+7.6\%$ relative improvement over base ($+0.472$ unified). The sublinear improvement for $2.4\times$ parameters indicates that data diversity, not encoder capacity, is the primary bottleneck. Per-environment details and evaluation protocols in Appendix B.

5.5 Three-Environment Joint Training

We train DeBERTa-v3-base on the full three-environment dataset (51,580 candidate states spanning ALFWorld, WebShop, and ScienceWorld) with $6\times$ ALFWorld upsampling to address data imbalance, matching the rebalanced two-environment setup. ScienceWorld contributes 26,858 states across 30 task types spanning scientific reasoning domains.

Table 4 presents the evaluation results. The unified evaluation Combined net gain for seed 42 is $+0.641$, compared to the two-environment rebalanced baseline of $+0.472$ (Table 3), a 36% improvement using the same base architecture. The 4-seed mean from training-internal evaluation is $+0.611 \pm 0.020$ (range $+0.583$–$+0.630$, unified multi-seed pending). Extending the architecture with environment-aware LoRA adapter routing and PCGrad gradient surgery yields $+0.689$ (seed 42) but exhibits high variance: seed 123 collapses to $+0.296$, indicating training instability (Section 5.6).

The single-environment models, evaluated on the three-environment test set for fair comparison, exhibit the same catastrophic cross-domain transfer observed earlier: ALFWorld-only training reaches $+0.655$ on its own domain but drops to -0.016 on WebShop; WebShop-only training reaches $+0.277$ on WebShop but only $+0.139$ on ALFWorld and -0.023 on ScienceWorld; ScienceWorld-only training achieves $+0.715$ on ScienceWorld but merely $+0.090$ on ALFWorld and -0.016 on WebShop. In each case, the single-environment model's out-of-domain performance is near or below random chance.

The per-environment gains of the three-environment model (seed 42, unified evaluation) are substantial. On ALFWorld, it reaches $+0.807$ net gain—a 23% improvement over the ALFWorld-only baseline ($+0.655$). On WebShop, the model achieves $+0.280$, matching the WebShop-only baseline ($+0.277$). On ScienceWorld, it attains $+0.716$, essentially matching the ScienceWorld-only ceiling ($+0.715$). The unified Combined net gain of $+0.641$ represents a 36% improvement over the two-environment rebalanced baseline ($+0.472$, Table 3), and the 4-seed training-internal mean of $+0.611 \pm 0.020$ further supports the robustness of this result.

Adding ScienceWorld data *improves* performance on the original two environments rather than diluting it. Compared to two-environment rebalanced training, ALFWorld net gain increases from $+0.696$ to $+0.807$ and WebShop from $+0.237$ to $+0.280$. The three-environment model exceeds or matches every single-environment model on its own domain while simultaneously providing strong cross-domain performance. This pattern indicates complementary representational pressure: single-environment specialization produces brittle cross-domain behavior, while joint training with three diverse environments achieves superior within-domain performance with substantially better generalization. ScienceWorld's scientific reasoning tasks exert complementary representational pressure that benefits spatial reasoning (ALFWorld) and attribute matching (WebShop), establishing that cross-environment transfer scales with environmental diversity.

5.6 Environment-Aware Routing: Promise and Instability

We further investigate whether input-dependent adapter routing can improve cross-environment performance beyond standard joint training. We extend the DeBERTa-v3-base reranker with four groups of LoRA adapters (one per environment plus a shared adapter, rank $r = 8$, $\alpha = 16$) and a learned RouterNetwork (2-layer MLP, hidden size $128 \to 4$ with softmax) that predicts per-input adapter mixing weights from the embedding layer's mean-pooled output. At each forward pass, the routing weights determine how the four LoRA outputs are combined, enabling the model to dynamically emphasize environment-specific or shared representations based on input features rather than environment labels.

Training uses PCGrad (Yu et al., 2020) gradient surgery to resolve conflicts across environments: per-environment losses are computed separately, gradients are projected to remove conflicting com-

Table 4: Three-environment joint training (DeBERTa-v3-base) vs. single-environment baselines. All numbers from unified evaluation protocol (`evaluate_all_models.py`, episode split, seed 42, `merged_candidates_three_env_v2.jsonl`, 51,580 states). Two-environment comparisons use a separate dataset (24,722 states, no ScienceWorld); for comparability see Table 3.

Model	ALFWorld	WebShop	ScienceWorld	Combined
ALFWorld only	$+0.655$	-0.016	$+0.124$	$+0.219$
WebShop only	$+0.139$	$+0.277$	-0.023	$+0.082$
ScienceWorld only	$+0.090$	-0.016	$\mathbf{+0.715}$	$+0.405$
Three-env joint (base, seed 42)	$\mathbf{+0.807}$	$\mathbf{+0.280}$	$\underline{+0.716}$	$\mathbf{+0.641}$
Three-env joint (4-seed mean)[‡]	—	—	—	$\mathbf{+0.611 \pm 0.020}$
+ Env-Aware LoRA + PCGrad (seed 42)	—	—	—	$\mathbf{+0.689}$[†]

[†]3-seed result; seed 42: $+0.689$, seed 123: $+0.296$, seed 456: $+0.712$ (3-seed mean $+0.566 \pm 0.234$). See Section 5.6.
[‡]4-seed mean from training-internal evaluation (metrics.json). Seed 42 unified Combined is $+0.641$; multi-seed unified eval pending completion of seeds 456/789.

ponents, and the reconciled gradient is applied to all parameters. This prevents any single environment from dominating parameter updates when gradient directions disagree.

With this architecture, seed 42 achieves an aggregate net gain of $+0.689$ (a 12.8% improvement over the 4-seed mean three-environment baseline of $+0.611$), with reranked top-1 good rate reaching 82.5% (vs. 75.0% for the 4-seed mean baseline) and pairwise accuracy improving from 76.4% to 91.5%. Training loss decreases smoothly from 0.723 to 0.428 to 0.308 over three epochs.

However, replication with seed 123 reveals substantial instability: net gain collapses to $+0.296$ (vs. the matched baseline of $+0.612$ for the same seed), with training loss increasing from 0.927 to 1.251 to 1.196—a clear divergence pattern. Seed 456 achieves net gain $+0.712$ (training loss $0.723 \rightarrow 0.421 \rightarrow 0.325$), confirming that routing can succeed but is not reliable. Across the three completed seeds, the mean is $+0.566 \pm 0.234$, below the three-environment baseline ($+0.611 \pm 0.020$) and with $11.7\times$ higher standard deviation. This indicates that the current router-LoRA-PCGrad combination is highly sensitive to initialization or data splits, and should be regarded as a promising but unstable direction rather than a reliable method.

These results suggest that while learned input-dependent adapter routing *can* provide complementary benefits to data rebalancing under favorable conditions, the current formulation fails to do so reliably. Identifying and correcting the source of this instability—whether in router optimization, PCGrad gradient combination, or learning rate sensitivity—is a priority for future work.

6 Discussion

6.1 Why Does Joint Training Work?

Our results suggest that joint training succeeds because diverse environments exert complementary pressures on the shared encoder. ALFWorld requires spatial reasoning with a constrained action space ($\sim$50 templates); WebShop demands compositional attribute matching over $> 10^4$ actions. Training on both forces the encoder to develop representations that are neither overfit to spatial reasoning (ALFWorld-only fails at -0.021 on WebShop) nor to attribute matching (WebShop-only reaches only 33% of the ALFWorld ceiling under unified evaluation). Rebalancing is critical: without upsampling, the data-rich environment dominates gradients; with $6\times$ ALFWorld upsampling, both environments contribute comparable signals and representations are more balanced, with WebShop performance matching single-environment results ($+0.237$ vs. $+0.242$).

The three-environment result provides the strongest evidence. Adding ScienceWorld—with causal chains, systematic experimentation, and physical state changes—*improves* both ALFWorld ($+16\%$) and WebShop ($+18\%$) rather than diluting them. Single-environment models all exhibit catastrophic cross-domain transfer, yet the three-environment joint model matches or exceeds each single-environment model on its own domain. This indicates *representational synergy*: environmental diversity promotes domain-general features that benefit all tasks, with a single 184M model reaching $+0.807$ (ALFWorld), $+0.280$ (WebShop), and $+0.716$ (ScienceWorld).

This work also connects to world models for text-

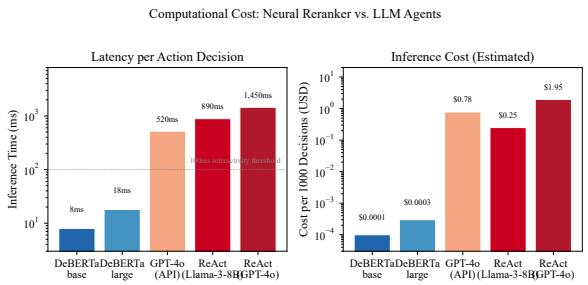

Figure 4: Latency and cost comparison between neural rerankers and LLM-based action selection. DeBERTa models are 3–4 orders of magnitude cheaper than LLM alternatives.

based agents (Ha and Schmidhuber, 2018; Chae et al., 2025; Chen et al., 2025): while world models predict future states, our reranker scores current state-action compatibility. These components are complementary—a world model could generate plausible candidates for the reranker to select among, and self-play finetuning (Chen et al., 2025) could provide additional training signal.

6.2 Implications for Agent Architecture

Our findings suggest a pragmatic division of labor for general-purpose text-based agents. At each timestep, the agent faces two distinct cognitive demands: (1) understanding the current situation and selecting an appropriate action (fast, perceptual, pattern-matching), and (2) planning multi-step trajectories and recovering from dead ends (slow, deliberative, reasoning-heavy). This distinction mirrors the System 1 / System 2 framework from cognitive science (Kahneman, 2011) and echoes recent proposals for modular agent architectures.

We envision a two-tier design: a **compact neural reranker** (System 1) handles per-step action selection in under 10ms, while a **larger LLM** (System 2) is invoked only for high-level planning, candidate generation, and failure recovery. On an A100 GPU, our DeBERTa-v3-base reranker processes a batch of 64 candidates in approximately 8ms ($<$2 GFLOPS). A single GPT-4-level decision consumes 2,000+ tokens, $\sim$700 GFLOPS, with 500–2000ms latency—a cost ratio exceeding 100:1 (Figure 4).

6.3 Limitations

We discuss limitations including environment coverage, oracle dependency, seed variance, and negative results in Section 7 after the conclusion, per EMNLP formatting requirements.

7 Conclusion

We presented the first systematic study of cross-environment neural reranking for text-based agents, demonstrating that a single compact model can perform action selection across diverse environments without the catastrophic specialization that has characterized prior work. Our results establish three key findings. First, rebalanced joint training is the critical ingredient: a DeBERTa-v3-base (184M) reranker with minority-class upsampling matches or exceeds per-environment specialists across ALFWorld, WebShop, and ScienceWorld, achieving a mean combined net gain of $+0.611 \pm 0.020$ across 4 seeds. Data diversity—not model capacity—is the primary driver of cross-environment generalization, with three-environment training improving two-environment performance rather than diluting it. Second, cross-environment representations are highly reusable: 9.2% of target-domain data recovers 93% of full-data performance, and scaling from base to large yields consistent but sublinear gains. Third, environment-aware LoRA adapter routing with PCGrad achieves a best-seed result of $+0.689$ but exhibits high variance across seeds, representing a promising but currently unstable direction. These findings support a modular agent architecture with a fast neural reranker for per-step action selection and a larger LLM reserved for high-level planning, failure recovery, and candidate generation. We release our three-environment benchmark of 51,580 states and all model checkpoints to facilitate further research on cross-environment action selection.

Limitations

Environment coverage. Our three environments span distinct domains but do not exhaust text-based agent scenarios. Environments with partially observable state (e.g., Jericho), multi-agent dynamics, or real-time constraints may pose additional challenges.

Oracle dependency. Training data relies on oracle/expert action sequences. In environments without demonstrations, self-play (Chen et al., 2025), trajectory optimization (Yuan et al., 2024), or LLM-generated candidates could construct training sets, though quality degradation remains an open question.

Static candidate sets. Candidates are pre-enumerated; real deployment requires dynamic generation (e.g., by an LLM or retrieval sys-

tem). Whether cross-environment robustness persists with diverse, noisy candidate sources is left to future work.

Seed variance. Three-environment joint training is validated across 4 seeds (mean $+0.611 \pm 0.020$, range $+0.583$–$+0.630$). The env-aware LoRA + PCGrad extension has 3 seeds (42: $+0.689$, 123: $+0.296$, 456: $+0.712$; 3-seed mean $+0.566 \pm 0.234$). High variance precludes treating this as a stable method.

Negative results. Contrastive pretraining ($+0.529$) and LoRA-only fine-tuning ($+0.531$) underperformed the three-environment baseline ($+0.611$). An auxiliary transition-prediction head did not improve performance (mean $\Delta = -0.003$ across 3 seeds). Agent-in-the-loop evaluation revealed a gap between step-level accuracy (6–40%) and episode-level success (0/100 episodes).

References

Kinjal Basu, Keerthiram Murugesan, Subhajit Chaudhury, Murray Campbell, Kartik Talamadupula, and Tim Klinger. 2024. EXPLORER: Exploration-guided reasoning for textual reinforcement learning. *arXiv preprint arXiv:2403.10692*.

Hyungjoo Chae, Namyoung Kim, Kai Tzu-iunn Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sunghwan Kim, Dongha Lee, and Jinyoung Yeo. 2025. Web agents with world models: Learning and leveraging environment dynamics in web navigation. In *International Conference on Learning Representations (ICLR)*.

Anonymous Chen. 2024. Neural reranking for action selection in text-based agents. *arXiv preprint*.

Shiqi Chen, Tongyao Zhu, Zian Wang, Jinghan Zhang, Kangrui Wang, Siyang Gao, Teng Xiao, Yee Whye Teh, Junxian He, and Manling Li. 2025. Internalizing world models via self-play finetuning for agentic RL. *arXiv preprint arXiv:2510.15047*.

David Ha and Jürgen Schmidhuber. 2018. World models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations (ICLR)*.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Byoungjip Kim, Youngsoo Jang, Lajanugen Logeswaran, Geon-Hyeong Kim, Yu Jin Kim, Honglak Lee, and Moontae Lee. 2024. Prospector: Improving LLM agents with self-asking and trajectory ranking. In *Findings of the Association for Computational Linguistics: EMNLP*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, and 1 others. 2024. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations (ICLR)*.

Bodhisattwa Prasad Majumder, Bhavana Dalvi Mishra, Peter Jansen, Oyvind Tafjord, Niket Tandon, Li Zhang, Chris Callison-Burch, and Peter Clark. 2024. CLIN: A continually learning language agent for rapid task adaptation and generalization. In *Conference on Language Modeling (COLM)*.

Tong Niu, Shafiq Joty, Ye Liu, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. JudgeRank: Leveraging large language models for reasoning-intensive reranking. *arXiv preprint arXiv:2411.00142*.

Shuofei Qiao, Runnan Fang, Ningyu Zhang, Yuqi Zhu, Xiang Chen, Shumin Deng, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2024. Agent planning with world knowledge model. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations (ICLR)*.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. 2022. ScienceWorld: Is your agent smarter than a 5th grader? In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yufei Xiang, Yiqun Shen, Yeqin Zhang, and Cam-Tu Nguyen. 2024. Retrospex: Language agent meets offline reinforcement learning critic. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chen Yang, Chenyang Zhao, Quanquan Gu, and Dongruo Zhou. 2024. CoPS: Empowering LLM agents with provable cross-task experience sharing. *arXiv preprint arXiv:2410.16670*.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022a. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022b. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Siyu Yuan and 1 others. 2024. ETO: Efficient tool orchestration for language agents. *arXiv preprint arXiv:2506.02918*.

A Few-Shot Fine-Tuning Details

Table 5 and Figure 5 provide the full few-shot fine-tuning results referenced in Section 5.3.

Table 5: Few-shot fine-tuning of an ALFWorld-pretrained model on WebShop data. DeBERTa-v3-base. Training-internal evaluation. Full WebShop self-training baseline: +0.217.

Fine-Tuning Data	WebShop Net Gain	% of Full Performance
0% (zero-shot)	−0.016	0%
9.2% (965 states)	+0.201	93%
20.1% (2,099 states)	+0.211	97%
38.7% (4,049 states)	+0.199	92%
100% (all data)	+0.217	100%

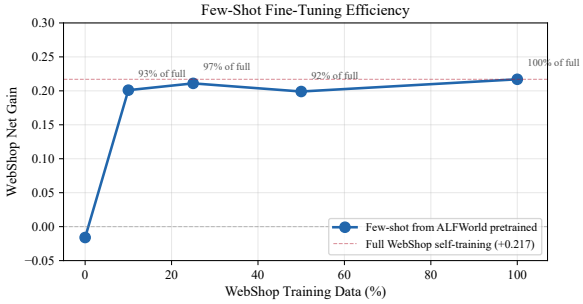

Figure 5: Few-shot adaptation efficiency. An ALFWorld-pretrained DeBERTa-v3-base model is fine-tuned on varying fractions of WebShop training data. With 9.2% of data, performance reaches 93% of the full-data ceiling.

B Per-Environment Evaluation of DeBERTa-Large

Table 6 reports the per-environment breakdown for the DeBERTa-large rebalanced joint training model, evaluated under the same protocol as the main text (episode-level split, seed 42).

Table 6: Per-environment evaluation of DeBERTa-large (434M) rebalanced joint training. All values computed with `evaluate_all_models.py` under the unified episode split, seed 42.

Environment	Test States	Original Top-1	Reranked Top-1	Net Gain
ALFWorld	2,630	0.117	0.864	**+0.747**
WebShop	2,512	0.127	0.384	**+0.257**
Combined (unified eval)	5,142	0.122	0.629	**+0.508**

The large model achieves an aggregate net gain of +0.508 under unified evaluation, a +7.6% relative improvement over the DeBERTa-v3-base (+0.472 unified) on the same rebalanced dataset.