Demonstration Selection for In-Context Learning via Reinforcement Learning

Xubin Wang, Jianfei Wu, Yichen Yuan, Deyu Cai, Mingzhe Li, Weijia Jia

BNU-BNBU Institute of Artificial Intelligence and Future Networks
Email: wangxubin@ieee.org
https://arxiv.org/pdf/2412.03966?

May 28, 2025

Outline

Introduction

- **LLM agents** represent a frontier in AI decision-making, yet face critical evolutionary limitations
- **Key gap**: Current approaches focus on *offline deployment*, causing:
  - Over-reliance on base model capabilities
  - Persistent misalignment issues
- **Core challenge**: Enable continuous evolution through environmental interaction
- **Critical barriers**:
  - Low sample efficiency
  - Poor interpretability
  - Resource constraints on edge devices
- **Our solution**: Novel reinforcement learning framework for *self-evolving LLM agents* with academic/practical significance

Introduction: Few-Shot Learning and ICL

- In few-shot learning, **In-Context Learning (ICL)** is a promising approach to enhance LLM reasoning.
- ICL uses LLMs (like GPT architecture) by providing a curated set of demonstrations as context, avoiding extensive retraining.
- This is suitable for tasks with limited labeled data.
- **Critical challenge:** The effectiveness of ICL depends on selecting appropriate and representative demonstrations.
- Careful selection influences generalization and accuracy in novel situations.

Introduction: Challenges in Demonstration Selection

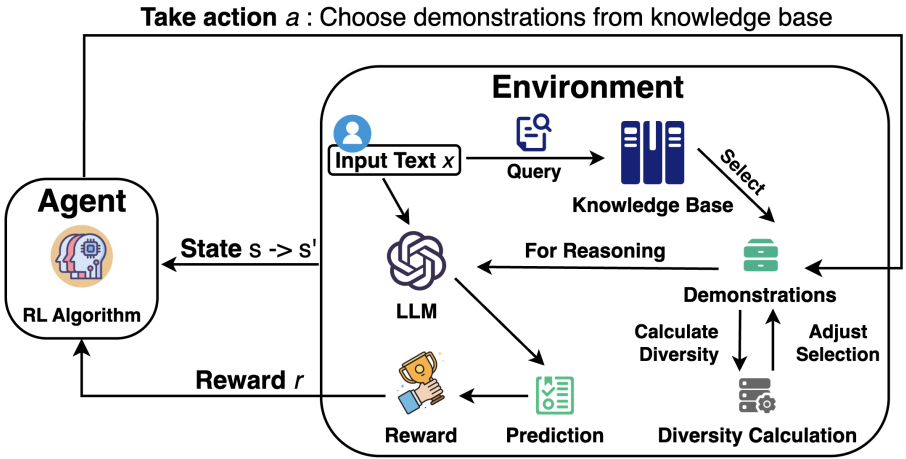
- Significant challenge in selecting the most relevant and diverse demonstrations from the knowledge base to optimize reasoning performance.
- Traditional methods often prioritize **similarity**, potentially overlooking diversity.
- This can lead to biased representations that don't generalize well to unseen data.
- Conventional techniques use fixed strategies, failing to dynamically adapt to task requirements.
- This rigidity limits ICL effectiveness, as selected demonstrations may not align optimally with task context or nuances.

An Example Shows How a Diversity-based Demonstration Method Works

Introduction: RDES & Contributions

- **Critical Gap:** Traditional ICL methods prioritize similarity-based demonstration selection, causing:
  - Limited generalization due to insufficient diversity
  - Suboptimal performance from static selection strategies
  - Biased representations in few-shot learning
- **RDES Solution:** Reinforcement learning framework for dynamic demonstration selection
  - Formulates selection as **sequential decision-making problem**
  - Jointly optimizes **relevance** (accuracy) and **diversity** (generalization)
  - Implements dual RL approaches: **Q-learning** and **PPO variant**
- **Key Contributions:**
  - Novel RL framework for adaptive demonstration selection
  - Optimization method mitigating overfitting while enhancing generalization
  - Seamless integration with **Chain-of-Thought (CoT)** reasoning
  - Comprehensive evaluation: **10 baselines** across **14 LLMs** showing significant gains

RDES Framework



Related Work: In-Context Learning (ICL)

- Transformative paradigm, especially with LLMs.
- Models adapt to new tasks by conditioning on a small set of demonstrations.
- Introduced with GPT-3, showing LLMs perform tasks with few exemplars.
- Effective in few-shot and zero-shot scenarios.
- **Critical dependency on demonstration selection**.
- Recent studies emphasize need for effective selection strategies.
- Approaches like DPP and IDS highlight diversity and relevance.

Related Work: Demonstration Selection Techniques

- Crucial for ICL success.
- Traditional: heuristics, statistical measures, informative examples (active learning concept).
- Recent prominence of **diversity**: improves generalization (e.g., clustering, coverage-based like BERTScore-Recall, representative sampling).
- Skill-based methods (Skill-KNN) optimize selection by eliminating irrelevant features.
- Some methods prioritize diversity statically (Yang et al., 2023).
- Some select based on uncertainty and diversity without training a policy (Mavromatis et al., 2023).
- Calibration techniques focus on correcting biases (Zhao et al., 2021).

Related Work: RL in Demonstration Selection

- Promising framework for optimizing selection, allowing policies to adapt iteratively based on performance feedback.
- Prior work: (Scarlatos & Lan, 2023) used RL for selection and sequencing; (Zhang et al., 2022) formulated selection as a sequential decision-making task using Q-learning.
- **RDES builds on these**: sits at intersection of ICL, selection, and RL-based post-training.
- **Distinction of RDES**: Uniquely focuses on the **dual objectives of diversity and relevance**, explicitly optimizing both via a diversity score in the reward function. 1) Explored Q-learning and PPO variants (RDES/PPO); 2) Integrates with CoT prompting (RDES/C); 3) Lighter-weight alternative to RLHF (optimizes input selection, not model weights); 4) Learns an adaptive policy per query, better optimizing selection than some static/fixed methods.

Methodology: Overview

- RDES tackles demonstration selection using a principled **RL approach**.
- Jointly optimizes for **relevance** and **diversity**.
- Four components:
  1. Formal problem formulation as a Markov Decision Process (MDP).
  2. Dual optimization strategies using Q-learning and PPO.
  3. Implementation details.
  4. Prompting strategies.

Methodology: RL Formulation (MDP)

- RL provides a natural framework for sequential decision making.
- Interaction modeled as iterative process: policy learns to construct optimal demonstration sets through trial-and-error.
- Formalized as a finite-horizon MDP $M = (S, A, P, R, \gamma)$.

Methodology: MDP Components - State Space (S)

- Captures complete decision context through four components:
  - **Textual features:** TF-IDF vector of input text $\phi_x(x_t)$.
  - **Demonstration memory:** Aggregated embeddings of selected examples $\phi_E(E_t)$.
  - **Prediction history:** One-hot encoded previous predictions $\phi_y(\hat{y}_t)$.
  - **Diversity tracking:** Normalized label diversity $D_t = |L(E_t)|/k$.
- State embedding: $\phi(s_t) = \phi_x(x_t) \oplus \phi_E(E_t) \oplus \phi_y(\hat{y}_t) \oplus \phi_D(D_t)$ (vector concatenation).

Methodology: MDP Components - Action Space (A) & Transition Dynamics (P)

- **Action Space (A):**
  - Discrete selection over candidate demonstrations K.
  - Action $a_t \in \{1, ..., |K|\}$ indicates the chosen example index from the knowledge base.
- **Transition Dynamics (P):**
  - Deterministic state updates through demonstration set modification.
  - Selecting candidate k_{a_t} in state s_t leads to next state s_{t+1}:
  - $s_{t+1} = f(s_t, a_t) = (x_t, E_t \cup \{k_{a_t}\}, \hat{y}_{t+1}, D_{t+1})$.
  - $\hat{y}_{t+1}$ is new prediction, D_{t+1} is new diversity score.

Methodology: MDP Components - Reward Function (R) & Discount Factor (γ)

- **Reward Function (R):**
  - Multi-objective reward balancing prediction accuracy and diversity gain.
  - $R(s_t, a_t) = \underbrace{I(y_{true} = \hat{y}_t)}_{\text{Accuracy}} + \lambda \underbrace{(D_{t+1} - D_t)}_{\text{Diversity Improvement}}$.
  - $I(\cdot)$ is indicator function, y_{true} is true label, $\hat{y}_t$ is prediction at step t.
  - D_t is diversity at t, D_{t+1} is diversity after adding example.
  - λ controls exploration-exploitation tradeoff.
  - λ adapts via annealing schedule: $\lambda(t) = \lambda_{min} + (\lambda_{max} - \lambda_{min})e^{-\eta t}$.
  - Schedule prioritizes early diversity exploration before focusing on accuracy.
- **Discount Factor (γ):**
  - $\gamma \in [0, 1)$ emphasizes immediate rewards.
  - Suitable for finite-horizon few-shot learning (fixed number of examples selected).

Methodology: Optimization Framework

- Two primary RL algorithms used to handle state space complexities.
- **Q-learning Approach:**
  - Model-free solution for learning strategies via temporal difference updates.
  - Effective for small/discretizable state spaces.
  - Action-value function $Q(s, a)$ estimates expected cumulative rewards.
  - Updates via standard Q-learning rule:
    $Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$.
  - Implementation aspects: state discretization (TF-IDF binning), ϵ-greedy exploration with exponential decay.
  - Uses tabular Q-value storage.
  - Theoretical convergence under standard conditions (Robbins-Monro, bounded rewards).

Methodology: Optimization Framework (cont.)

- **Proximal Policy Optimization (PPO) Variant:**
  - For high-dimensional state spaces where tabular methods are infeasible.
  - Actor-critic architecture using neural networks.
  - **Policy Network (π_θ):** Neural network producing demonstration selection probabilities $\pi_\theta(a|s)$. Uses state embedding $\phi(s)$.
  - **Value Network (V_ψ):** Neural network estimating state value $V_\psi(s)$ (expected cumulative reward). Uses state embedding $\phi(s)$.
  - **Optimization Objective:** PPO optimizes a **clipped surrogate objective** for stability.
  - Combines clipped surrogate loss (L^{CLIP}), value function loss (L^{VF}), and entropy bonus (L^{ENT}): $L(\theta, \psi) = E_t[L_t^{CLIP}(\theta) - c_1 L_t^{VF}(\psi) + c_2 L_t^{ENT}(\theta)]$.
  - L^{CLIP} uses probability ratio $r_t(\theta)$ and advantage estimate A_t, clipped within $[1-\epsilon, 1+\epsilon]$.
  - L^{VF} is squared error between predicted value and estimated return $\hat{R}_t$.
  - L^{ENT} encourages exploration.

Methodology: Algorithmic Implementation

- **Unified Training Paradigm:** Core procedure shared by Q-learning and PPO (Algorithm 1).
- Iteratively: Sample input, select demonstrations (initially based on relevance, adjusted for diversity), format prompt, get LLM prediction, compute diversity score, encode state, select action (example index), calculate reward (accuracy + diversity change), update policy parameters.
- **State Representation Details:** State embedding $\phi(s_t)$ is concatenation of TF-IDF vector, aggregated selected example embeddings, prediction history, and normalized label diversity. Provides context for selection decisions.

Methodology: RDES Training Framework

Algorithm 1 RDES Training Framework

Require: Knowledge base $\mathcal{K}$, Test inputs $\mathcal{D}_{\text{test}}$, LLM $\mathcal{M}$, RL algorithm $\mathcal{A}$

1: Initialize selection policy π (Q-table or neural networks)
2: Precompute TF-IDF vectors for each sample $x_i \in \mathcal{D}_{\text{test}}$ and for knowledge base $\mathcal{K}$
3: **for** $i = 1$ to N **do**
4: Sample input $x_i \in \mathcal{D}_{\text{test}}$
5: Select demonstrations E with initial candidates based on relevance (e.g., top-k TF-IDF matches from $\mathcal{K}$) and apply diversity adjustment
6: Generate prompt $p = \text{Format}(x_i, E)$
7: Obtain prediction $\hat{y} = \mathcal{M}(p)$
8: Compute diversity score $D = \frac{|\mathcal{L}(E)|}{k}$
9: Encode state $s = \phi(x_i, E, \hat{y}, D)$
10: Select action $a \sim \pi(s)$ (example index from $\mathcal{K}$)
11: Calculate reward $r = \mathbb{I}(y_{\text{true}} = \hat{y}) + \lambda(D_{\text{new}} - D_{\text{old}})$
12: Update policy parameters θ using $\mathcal{A}$ with (s, a, r)
13: **end for**
14: **return** Optimized policy π^*

Methodology: Prompting Strategies

- Used with selected examples to enhance LLM performance.
- **Standard Prompting:**
  - Concatenate input text, selected demonstrations (input-output pairs), and possible labels.
  - LLM predicts label probability given this prompt structure.
  - $p(y|x, E) = \text{PLM}(y \mid \text{Prompt} : x, \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^{k}, Y)$.
- **CoT Prompting:**
  - Incorporates CoT reasoning steps before final label.
  - LLM generates intermediate reasoning steps.
  - Formulated as marginalizing over possible reasoning chains R.
  - $p(y|x, E) = \sum_{r \in R} \text{PLM}(r|x, E) \cdot \text{PLM}(y|x, E, r)$.
  - Model computes probability of reasoning chain, then label probability conditioned on input, demos, and reasoning chain.

Experiments: Datasets

- **Four main classification datasets:**
  - **BANKING77:** Banking sector intents. (77 intents, 9003 KB, 3080 Test, 1 Domain).
  - **HWU64:** Extensive multi-domain coverage. (64 intents, 8828 KB, 1104 Test, 21 Domains).
  - **LIU54:** Extensive multi-domain coverage, specialized queries. (54 intents, 20382 KB, 2548 Test, 21 Domains).
  - **CLINC150:** Further enriches evaluation, technical nature. (150 intents, 18000 KB, 2250 Test, 10 Domains).
- **Additional challenging reasoning benchmarks**:
  - BigBenchHard (boolean expressions, web of lies subsets).
  - GSM-8K (math word problems).
  - SST5 (sentiment treebank).
- Randomly sampled 1,000 examples from test sets for supplementary experiments.

Experiments: Compared Methods

- Evaluated ten baseline approaches.
- **Prompt Engineering Methods:**
  - Zero-Shot (ZS): No demonstrations, tests generalization.
  - Knowledge Prompting (KP): Contextual information generation.
  - Least-to-Most (L2M): Break tasks into smaller steps.
  - Chain of Thought (CoT): Step-by-step reasoning.
  - Self-Refine (SF): Iterative critique and refinement.
- **Demonstration Selection Methods:**
  - Few-Shot (FS): Limited text-label pairs.
  - Few-Shot with CoT (FSC): FS + explanations.
  - Active Selection (AES): Iterative selection using RL.
  - Representative Selection (RDS): Diverse subset identification.
  - Adaptive Selection (ADA): Uncertainty-based, semantic diversity.
- **Ours:** RDES/B (base), RDES/C (RDES/B + CoT), RDES/PPO (PPO variant).

Experiments: LLMs Used

- Diverse set, including closed-source and open-source models.
- **Closed-source** (proprietary, strong NLP capabilities):
  - GPT-3.5-turbo (OpenAI).
  - Doubao-lite-4k, Doubao-pro-4k (ByteDance).
  - Hunyuan-lite (Tencent).
- **Open-source** (flexibility, customization, innovation):
  - Gemma-2-2B, Gemma-2-9B (Google).
  - LLaMA-3.2-1B, LLaMA-3.2-3B, LLaMA-3-8B (Meta).
  - Qwen-2.5-7B, Qwen-2.5-14B, Qwen-1.5-72B (Alibaba Cloud).
- **Specific models for challenging tasks** (Section 4.5, Appendix A.6):
  - Qwen-2.5-72B.
  - DeepSeek-R1-32B (focused on reasoning).

Results: Reasoning Performance (Closed-Source)

Table: Performance comparison of methods designed to boost LLM reasoning across various datasets on closed-source LLMs, with a focus on accuracy.

Datasets	Models	Prompt Engineering Methods					Demonstration Selection Methods					Ours	
		ZS	KP	L2M	CoT	SF	FS	FSC	AES	RDS	ADA	RDES/B	RDES/C
BANKING77	GPT-3.5-turbo	0.340	0.240	0.260	0.200	0.380	0.520	0.320	0.260	0.240	0.360	0.767	**0.858**
	Doubao-lite-4k	0.300	0.300	0.300	0.320	0.300	0.500	0.360	0.300	0.280	0.400	0.750	**0.830**
	Doubao-pro-4k	0.500	0.400	0.500	0.480	0.600	0.540	0.540	0.700	0.680	**0.900**	0.838	0.888
	Hunyuan-lite	0.300	0.233	0.433	0.200	0.300	0.233	0.133	0.320	0.320	0.600	0.593	**0.775**
	Average	0.360	0.293	0.373	0.300	0.395	0.448	0.338	0.395	0.380	0.565	0.737	**0.838**
CLINC150	GPT-3.5-turbo	0.460	0.420	0.400	0.480	0.460	0.600	0.380	0.300	0.380	0.720	0.845	**0.949**
	Doubao-lite-4k	0.700	0.600	0.600	0.700	0.500	0.680	0.440	0.680	0.660	0.700	0.825	**0.927**
	Doubao-pro-4k	0.660	0.680	0.620	0.700	0.700	0.800	0.640	0.680	0.640	0.900	0.938	**0.961**
	Hunyuan-lite	0.633	**0.800**	0.767	0.700	0.633	0.467	0.500	0.480	0.620	**0.800**	0.730	0.772
	Average	0.613	0.625	0.597	0.645	0.573	0.637	0.490	0.535	0.575	0.780	0.835	**0.902**
HWU64	GPT-3.5-turbo	0.260	0.360	0.280	0.340	0.280	0.560	0.360	0.100	0.260	0.520	0.850	**0.914**
	Doubao-lite-4k	0.500	0.500	0.500	0.480	0.500	0.520	0.340	0.360	0.420	0.700	0.765	**0.873**
	Doubao-pro-4k	0.640	0.760	0.620	0.800	0.640	0.680	0.600	0.620	0.640	**1.000**	0.862	0.918
	Hunyuan-lite	0.533	0.367	0.333	0.433	0.233	0.600	0.433	0.540	0.320	0.700	0.514	**0.784**
	Average	0.483	0.497	0.433	0.513	0.413	0.590	0.433	0.405	0.410	0.730	0.748	**0.872**
LIU54	GPT-3.5-turbo	0.380	0.260	0.360	0.460	0.240	0.480	0.480	0.140	0.180	0.300	0.743	**0.868**
	Doubao-lite-4k	0.500	0.400	0.500	0.540	0.660	0.600	0.440	0.520	0.520	0.600	0.690	**0.841**
	Doubao-pro-4k	0.400	0.420	0.400	0.520	0.520	0.800	0.760	0.500	0.520	**0.900**	0.829	0.884
	Hunyuan-lite	0.533	0.500	0.567	0.700	0.633	0.367	0.500	0.460	0.620	0.560	0.565	**0.704**
	Average	0.453	0.395	0.457	0.555	0.513	0.562	0.545	0.405	0.460	0.590	0.707	**0.824**

Results: Reasoning Performance (Closed-Source) (cont.)

- RDES/B and RDES/C consistently outperform alternative methods across evaluated datasets (BANKING77, CLINC150, HWU64, LIU54).
- **RDES/C (with CoT) achieves highest accuracy in nearly all instances**.
- CoT and KP yield strong results, but task/dataset dependent.
- ADA and FSC competitive but generally outperformed by RDES/B and RDES/C.
- Doubao-pro-4k excelled, achieving peak performance of 0.961 on CLINC150 with RDES/C.
- GPT-3.5-turbo shows stable performance.
- RDES/C's incorporation of CoT consistently leads to superior performance.
- Underscores impact of advanced techniques, adaptive/CoT strategies are essential.

Results: Reasoning Performance (Open-Source)

Datasets	Models	Prompt Engineering Methods					Demonstration Selection Methods					Ours	
		ZS	KP	L2M	CoT	SF	FS	FSC	AES	RDS	ADA	RDES/B	RDES/C
BANKING77	Gemma-2-2B	0.200	0.280	0.200	0.200	0.260	0.300	0.340	0.280	0.220	**0.900**	0.831	0.861
	Gemma-2-9B	0.560	0.400	0.500	0.500	0.400	0.440	0.380	0.400	0.400	0.700	0.831	**0.886**
	LLaMA-3.2-1B	0.120	0.100	0.000	0.000	0.100	0.000	0.000	0.020	0.040	0.680	0.024	**0.744**
	LLaMA-3.2-3B	0.200	0.200	0.400	0.500	0.300	0.320	0.060	0.360	0.440	0.700	0.770	**0.805**
	LLaMA-3-8B	0.578	0.560	0.563	0.552	0.458	0.090	0.182	0.531	0.536	0.758	0.784	**0.847**
	Qwen-2.5-7B	0.700	0.480	0.600	0.420	0.480	0.440	0.180	0.440	0.420	0.700	0.803	**0.859**
	Qwen-2.5-14B	0.400	0.400	0.420	0.420	0.420	0.480	0.460	0.500	0.520	0.800	0.839	**0.868**
	Qwen-1.5-72B	0.529	0.480	0.524	0.528	0.551	0.653	0.612	0.509	0.542	0.775	0.785	**0.892**
	Average	0.411	0.363	0.401	0.390	0.371	0.340	0.277	0.380	0.390	0.752	0.708	**0.845**
CLINC150	Gemma-2-2B	0.400	0.600	0.420	0.460	0.560	0.560	0.540	0.500	0.380	0.800	0.875	**0.929**
	Gemma-2-9B	0.700	0.700	0.800	0.800	0.700	0.800	0.680	0.800	0.800	0.780	**0.864**	0.819
	LLaMA-3.2-1B	0.400	0.520	0.060	0.380	**0.600**	0.000	0.020	0.080	0.080	0.400	0.256	0.122
	LLaMA-3.2-3B	0.800	0.700	0.800	0.400	0.700	0.580	0.260	0.600	0.680	0.800	**0.845**	0.703
	LLaMA3-8B	0.523	0.439	0.594	0.504	0.569	0.007	0.285	0.571	0.543	0.767	**0.840**	0.783
	Qwen-2.5-7B	0.740	0.800	0.800	0.780	0.700	0.740	0.460	0.780	0.780	0.800	**0.879**	0.741
	Qwen-2.5-14B	0.900	0.900	0.900	0.800	0.840	0.700	0.700	0.740	0.740	0.900	**0.944**	0.792
	Qwen-1.5-72B	0.726	0.517	0.683	0.641	0.660	0.850	0.652	0.696	0.656	0.861	0.897	**0.963**
	Average	0.649	0.647	0.632	0.596	0.666	0.530	0.450	0.596	0.582	0.763	**0.800**	0.731
HWU64	Gemma2-2B	0.300	0.320	0.300	0.300	0.400	0.460	0.440	0.420	0.360	0.600	0.832	**0.851**
	Gemma2-9B	0.600	0.600	0.600	0.600	0.600	0.700	0.700	0.700	0.700	0.800	0.877	**0.910**
	LLaMA-3.2-1B	0.200	0.100	0.080	0.000	0.100	0.020	0.000	0.020	0.060	0.360	0.381	**0.687**
	LLaMA-3.2-3B	0.300	0.100	0.300	0.200	0.300	0.220	0.180	0.300	0.300	0.700	0.747	**0.817**
	LLaMA-3-8B	0.478	0.407	0.493	0.479	0.563	0.632	0.498	0.651	0.645	0.837	0.816	**0.859**
	Qwen-2.5-7B	0.780	0.700	0.800	0.600	0.800	0.640	0.540	0.760	0.740	0.800	0.805	**0.880**
	Qwen-2.5-14B	0.780	0.800	0.800	0.440	0.740	0.800	0.800	0.720	0.680	**0.900**	0.886	0.895
	Qwen-1.5-72B	0.698	0.615	0.676	0.661	0.668	0.825	0.817	0.749	0.774	0.877	0.867	**0.924**
	Average	0.517	0.455	0.506	0.410	0.521	0.537	0.497	0.540	0.532	0.734	0.776	**0.853**
LIU54	Gemma-2-2B	0.400	0.400	0.500	0.400	0.400	0.620	0.480	0.500	0.440	0.600	0.733	**0.854**
	Gemma-2-9B	0.500	0.500	0.600	0.600	0.600	0.580	0.580	0.500	0.500	**1.000**	0.722	0.837
	LLaMA-3.2-1B	0.200	0.160	0.300	0.400	0.080	0.040	0.040	0.360	0.320	**0.700**	0.058	0.651
	LLaMA-3.2-3B	0.400	0.400	0.400	0.300	0.400	0.360	0.320	0.500	0.400	0.600	**0.772**	0.749
	LLaMA-3-8B	0.358	0.409	0.428	0.360	0.392	0.396	0.320	0.347	0.312	0.763	0.779	**0.811**
	Qwen-2.5-7B	**0.800**	0.700	0.700	0.640	0.520	0.620	0.500	0.500	0.660	**0.800**	0.794	0.765
	Qwen-2.5-14B	0.700	0.860	0.700	0.660	0.700	0.660	0.600	0.740	0.780	**1.000**	0.849	0.743
	Qwen-1.5-72B	0.496	0.445	0.487	0.491	0.550	0.609	0.647	0.514	0.492	0.769	0.781	**0.880**
	Average	0.482	0.484	0.514	0.481	0.455	0.486	0.436	0.495	0.488	0.779	0.686	**0.786**

Results: Reasoning Performance (Open-Source)(cont.)

- Significant performance variations across datasets/models.
- **RDES/C consistently outperforms other methods in BANKING77** (average 0.845 vs ADA 0.752) and shows robustness in HWU64 (average 0.853 vs ADA 0.734).
- CLINC150 benefits from larger models (Qwen-1.5-72B).
- **On CLINC150, RDES/B (0.800) outperforms RDES/C (0.731)**, and ADA (0.763). This suggests dataset characteristics influence optimal RDES variant.
- ZS and KP show limitations compared to ADA and RDES.
- Larger models (Qwen-2.5-14B, Qwen-1.5-72B) show marked improvements, especially with RDES/C (synergistic effect of scale and technique).
- RDES methods, particularly with CoT, provide advantage across datasets.
- Dataset-specific trends underscore importance of tailored approaches.

Results: Average Performance

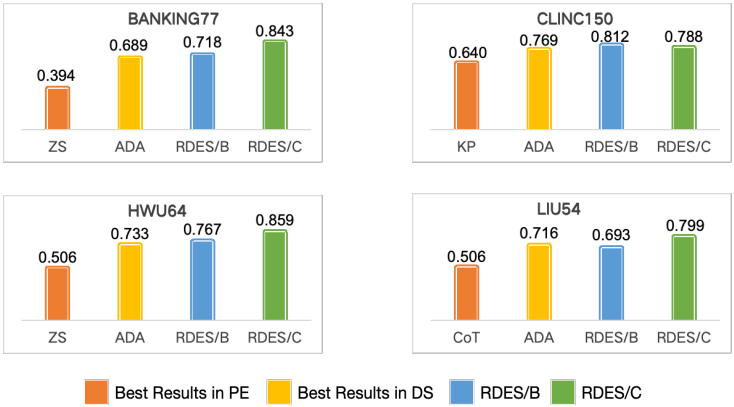


Figure: These figures illustrate the average results across closed-source/open-source models on different datasets, comparing the best results from the prompt engineering (PE) and demonstration selection (DS) methods with our proposed approach.

Results: Average Performance (cont.)

- Summarizes average performance across closed-source/open-source models.
- Highlights effectiveness of RDES/B and RDES/C compared to baselines.
- **BANKING77:** RDES/C (0.843) significantly surpasses RDES/B (0.718) and ADA (0.689).
- **CLINC150:** RDES/B (0.812) strong, followed by RDES/C (0.788), ADA (0.769).
- **HWU64:** RDES/C (0.859) leads, RDES/B (0.767), ADA (0.733).
- **LIU54:** RDES/C (0.799) outperforms RDES/B (0.693) and ADA (0.716).
- Overall illustrates effectiveness in enhancing performance through advanced techniques.

Results: Challenging Reasoning Tasks

- **RDES shows competitive performance**.
- Findings demonstrate RDES maintains strong performance on tasks requiring complex reasoning.
- Supports RDES's broader applicability beyond straightforward classification.
- Exploration of PPO shows promise for these tasks.

Table: Supplementary Performance Comparison on SST5, BigBenchHard, and GSM-8K

Methods	SST5		BigBenchHard - boolean expressions		BigBenchHard - web of lie		GSM-8K	
	Qwen-2.5-72B	DeepSeek-R1-32B	Qwen-2.5-72B	DeepSeek-R1-32B	Qwen-2.5-72B	DeepSeek-R1-32B	Qwen-2.5-72B	DeepSeek-R1-32B
FS	0.56	0.70	0.98	0.38	0.58	0.98	0.50	0.28
FSC	0.54	0.66	0.60	0.46	1.00	1.00	0.56	0.64
AES	0.84	0.84	0.53	0.60	0.85	0.72	0.92	0.08
RDS	0.76	0.84	0.53	0.60	0.89	0.68	0.90	0.48
ADA	0.90	0.90	0.53	0.60	0.83	0.72	0.98	0.36
RDES/B	0.44	0.57	0.76	1.00	0.50	0.93	0.87	0.37
RDES/C	0.51	0.52	0.90	0.99	0.98	1.00	0.92	0.73
RDES/PPO	0.84	0.84	1.00	1.00	1.00	0.90	0.94	0.48

Results: Varying Number of Demonstrations

- Investigates impact of varying number of demonstrations (k) on GSM-8K and SST5.
- Uses Qwen-2.5-72B model.
- Methods: FS, FSC, AES, RDS, ADA, RDES/B, RDES/C, RDES/PPO evaluated for $k = 3, 5, 7, 10$.
- Performance changes depending on the size of the demonstration set.
- Example GSM-8K: AES, RDS, ADA, RDES/PPO, RDES/C perform strongly at $k = 3, 7$, but performance can drop significantly at $k = 5, 10$ for some methods (e.g., AES, ADA).
- RDES/C seems more stable across k on GSM-8K compared to some baselines.
- SST5 performance is less sensitive to k variations for most methods.
- Results highlight diversity-driven generalization, especially with RDES/PPO, even in varying settings.

Results: Varying Number of Demonstrations (cont.)

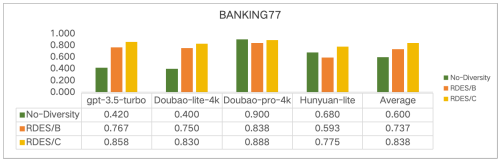
Table: Performance of Methods Across Varying Numbers of Demonstrations (k) Using Qwen-2.5-72B Model

Methods	GSM-8K (Accuracy)				SST5 (Accuracy)			
	k=3	k=5	k=7	k=10	k=3	k=5	k=7	k=10
FS	0.50	0.28	0.50	0.28	0.54	0.56	0.54	0.56
FSC	0.56	0.64	0.56	0.64	0.52	0.54	0.52	0.54
AES	0.92	0.08	0.92	0.08	0.82	0.84	0.82	0.84
RDS	0.90	0.48	0.90	0.48	0.74	0.76	0.74	0.76
ADA	0.98	0.36	0.98	0.36	0.88	0.90	0.88	0.90
RDES/B	0.87	0.37	0.87	0.37	0.42	0.44	0.42	0.44
RDES/C	0.92	0.73	0.92	0.73	0.49	0.51	0.49	0.51
RDES/PPO	0.94	0.48	0.94	0.48	0.82	0.84	0.82	0.84

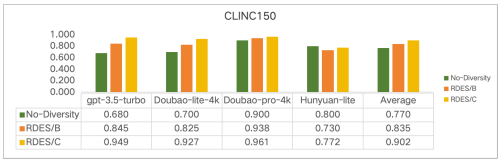
Results: Ablation Study on Diversity

- **Study Focus:** Impact of diversity mechanisms (No-Diversity, RDES/B, RDES/C) across closed-source and open-source models on four datasets.
- **Key Finding:** Incorporating diversity generally enhances model performance.
- **Closed-source Models (Figure 2):**
  - RDES/C consistently outperforms others across all datasets.
  - Example: BANKING77 avg accuracy: RDES/C (0.838) vs. No-Diversity (0.600).
- **Open-source Models (Figure 3):**
  - Performance varies by dataset.
  - BANKING77: RDES/C (0.845) vs. No-Diversity (0.747).
  - CLINC150: RDES/B (0.800) $>$ No-Diversity (0.768) and RDES/C (0.731).
  - HWU64: RDES/C (0.853) significantly boosts accuracy from No-Diversity (0.732).
  - LIU54: RDES/C (0.786) slightly higher than No-Diversity/RDES/B.
- **Conclusion:** A nuanced approach is needed for model and dataset pairing in open-source models.

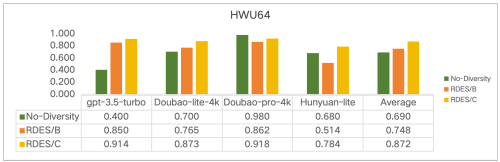
Results: Ablation Study on Diversity (Closed-source Models)



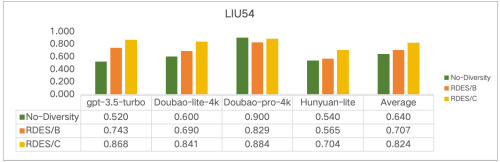
	gpt-3.5-turbo	Doubao-lite-4k	Doubao-pro-4k	Hunyuan-lite	Average
No-Diversity	0.420	0.400	0.900	0.680	0.600
RDES/B	0.767	0.750	0.838	0.593	0.737
RDES/C	0.858	0.830	0.888	0.775	0.838



	gpt-3.5-turbo	Doubao-lite-4k	Doubao-pro-4k	Hunyuan-lite	Average
No-Diversity	0.680	0.700	0.900	0.800	0.770
RDES/B	0.845	0.825	0.938	0.730	0.835
RDES/C	0.949	0.927	0.961	0.772	0.902



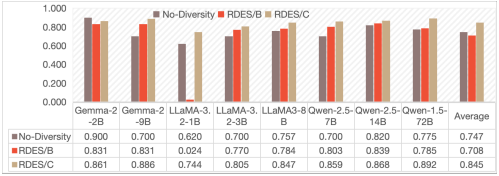
	gpt-3.5-turbo	Doubao-lite-4k	Doubao-pro-4k	Hunyuan-lite	Average
No-Diversity	0.400	0.700	0.980	0.680	0.690
RDES/B	0.850	0.765	0.862	0.514	0.748
RDES/C	0.914	0.873	0.918	0.784	0.872



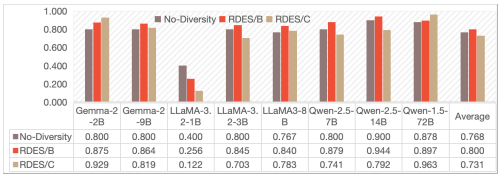
	gpt-3.5-turbo	Doubao-lite-4k	Doubao-pro-4k	Hunyuan-lite	Average
No-Diversity	0.520	0.600	0.900	0.540	0.640
RDES/B	0.743	0.690	0.829	0.565	0.707
RDES/C	0.868	0.841	0.884	0.704	0.824

Figure: Performance of various closed-source models across different datasets, highlighting the impact of diversity mechanisms.

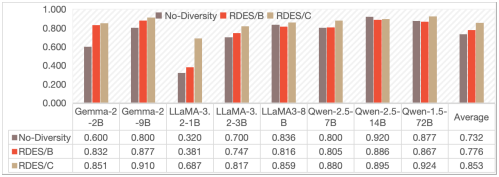
Results: Ablation Study on Diversity (Open-source Models)



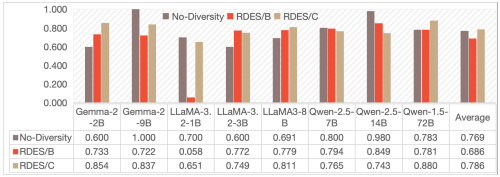
	Gemma-2-2B	Gemma-2-9B	LLaMA-3.2-1B	LLaMA-3.2-3B	LLaMA3-8B	Qwen-2.5-7B	Qwen-2.5-14B	Qwen-1.5-72B	Average
No-Diversity	0.900	0.700	0.620	0.700	0.757	0.700	0.820	0.775	0.747
RDES/B	0.831	0.831	0.024	0.770	0.784	0.803	0.839	0.785	0.708
RDES/C	0.861	0.886	0.744	0.805	0.847	0.859	0.868	0.892	0.845



	Gemma-2-2B	Gemma-2-9B	LLaMA-3.2-1B	LLaMA-3.2-3B	LLaMA3-8B	Qwen-2.5-7B	Qwen-2.5-14B	Qwen-1.5-72B	Average
No-Diversity	0.800	0.800	0.400	0.800	0.767	0.800	0.900	0.878	0.768
RDES/B	0.875	0.864	0.256	0.845	0.840	0.879	0.944	0.897	0.800
RDES/C	0.929	0.819	0.122	0.703	0.783	0.741	0.792	0.963	0.731



	Gemma-2-2B	Gemma-2-9B	LLaMA-3.2-1B	LLaMA-3.2-3B	LLaMA3-8B	Qwen-2.5-7B	Qwen-2.5-14B	Qwen-1.5-72B	Average
No-Diversity	0.600	0.800	0.320	0.700	0.836	0.800	0.920	0.877	0.732
RDES/B	0.832	0.877	0.381	0.747	0.816	0.805	0.886	0.867	0.776
RDES/C	0.851	0.910	0.687	0.817	0.859	0.880	0.895	0.924	0.853



	Gemma-2-2B	Gemma-2-9B	LLaMA-3.2-1B	LLaMA-3.2-3B	LLaMA3-8B	Qwen-2.5-7B	Qwen-2.5-14B	Qwen-1.5-72B	Average
No-Diversity	0.600	1.000	0.700	0.600	0.691	0.800	0.980	0.783	0.769
RDES/B	0.733	0.722	0.058	0.772	0.779	0.794	0.849	0.781	0.686
RDES/C	0.854	0.837	0.651	0.749	0.811	0.765	0.743	0.880	0.786

Figure: Performance of various open-source models across different datasets, highlighting the impact of diversity mechanisms.

Conclusion

- Introduced **RDES**, a novel framework using RL (Q-learning, with PPO variant) to optimize demonstration selection for ICL in LLMs.
- RDES balances **relevance and diversity**, enhancing generalization and mitigating overfitting.
- Extensive evaluation showed RDES **significantly outperforms ten baseline methods** on four benchmark classification datasets.
- Integrating RDES with **CoT reasoning (RDES/C) generally enhances performance**, though benefits vary with model and dataset.
- Additional experiments on more **challenging reasoning benchmarks** and with varying numbers of demonstrations further **validated RDES's effectiveness**.
- Highlighted **diversity-driven generalization**, especially with the RDES/PPO variant, even in complex tasks or varying settings.
- Results underscore the **potential of RL** to facilitate adaptive demonstration selection and address NLP complexities.

Future Work

- Refining diversity metrics.
- Extending RDES to tasks beyond classification (e.g., generation, question answering).
- Making CoT usage adaptive within the RL framework.
- Analyzing computational cost and sample efficiency.
- Exploring different retrieval methods.
- Assessing the generalization capabilities of strategies across datasets.

Impact Statement

- **Primary Positive Impact:** Significant enhancement in LLM accuracy and robustness in data-limited scenarios.
- Makes LLMs more effective for practical applications (intent detection, sentiment analysis).
- Helps mitigate overfitting biases from purely similarity-based selection.
- **Potential Negative Impacts:**
  - Enhanced classification could be misused (surveillance, censorship).
  - Extending to generative tasks could contribute to misinformation.
  - Training involves significant computational cost (numerous LLM calls), potentially limiting accessibility.
  - Lack of user studies means real-world human-centric impacts not yet.
- **Mitigation:** Explore computational efficiency, necessitate safeguards (especially for generation), conduct user studies, adherence to ethical principles (transparency, fairness, accountability).

Thank you for your attention!

Questions?