Demonstration Selection for In-Context Learning via Reinforcement Learning

RDES: Relevance-Diversity Enhanced Selection

Talk Info

Speaker	Xubin Wang
Paper	RDES (ICML 2025)
Flow	Problem $\rightarrow$ Method $\rightarrow$ Evidence
Contact	wangxubin@ieee.org

The ICL Demonstration Selection Problem

In-Context Learning (ICL)

- A few demonstrations in the prompt —no weight updates.
- Performance can swing wildly depending on which demonstrations you choose.

Why It Matters

- Real tasks (intent detection, classification, reasoning) with scarce, expensive labeled data.
- No budget or time for fine-tuning; LLM access is black-box only.

Core Challenge

Select a small set of demonstrations that is both **relevant** to the query and **diverse** in labels and structures.

Motivation: Similarity-Only Selection Fails

The Problem with Pure Similarity

Similarity-based retrieval keeps picking the same majority-class or near-duplicate examples. This gives poor coverage of minority intents and hard boundary cases.

Diversity is not just a regularizer —it is a core objective that directly improves robustness and generalization.

Why Existing Methods Fall Short

Common Approaches

- Similarity / embedding KNN
- Uncertainty sampling
- Static diversity (DPP, clustering)
- Heuristic prompt engineering

What They Miss

- No **query-adaptive policy**
- No joint relevance + diversity optimization
- One-shot selection (no correction)
- Ignore actual LLM feedback

Our approach: Keep the LLM completely frozen. We learn only a lightweight policy that decides which demonstrations to put in the prompt —no expensive weight updates like RLHF.

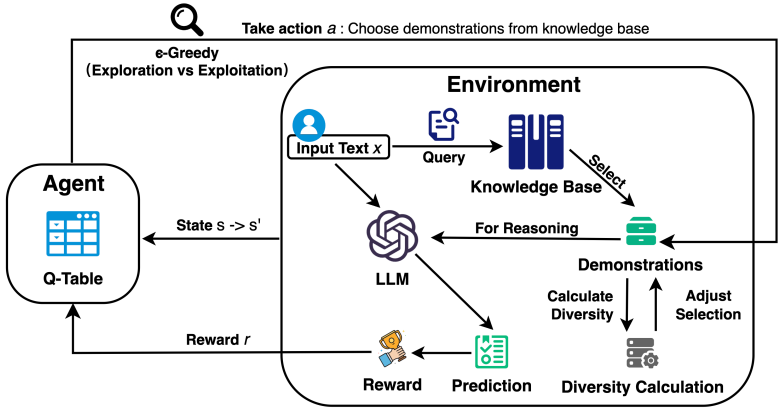
RDES Core Idea

We treat demonstration selection as a sequential decision process.

Relevance **Diversity** **Adaptivity**

Goal: For every query, dynamically build a small set that jointly maximizes accuracy and coverage.

RDES Framework: Agent—Environment Loop

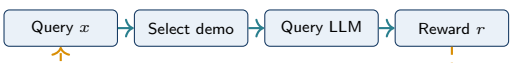


The agent sequentially selects demonstrations from a labeled pool, queries the frozen LLM, receives a reward (accuracy + diversity gain), and updates its policy. The final selected set is used for ICL.

MDP Formulation

MDP Components

$$
\begin{aligned}
s_t &= \phi_x(x) \oplus \phi_E(E_t) \oplus \phi_y(\hat{y}_t) \oplus D_t \\
a_t &\in \{1, \ldots, |\mathcal{K}|\} \\
s_{t+1} &= (x,\ E_t \cup \{k_{a_t}\},\ \hat{y}_{t+1},\ D_{t+1})
\end{aligned}
$$



State Design (4 parts)

- What the current query is about (simple TF-IDF features)
- Which demonstrations have already been picked
- The LLM's predictions so far
- How diverse the labels are in the current set

Sequential decisions let the policy correct early mistakes —impossible in one-shot retrieval.

Reward Design: Accuracy + Annealed Diversity

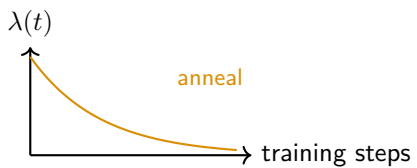
Reward

$$r_t = \underbrace{\mathbb{I}(y = \hat{y}_{t+1})}_{\text{Accuracy}} + \lambda(t) \underbrace{(D_{t+1} - D_t)}_{\Delta\text{Diversity}}$$

Annealing Schedule

$\lambda(t) = \lambda_{\max} \cdot e^{-\eta t}$ (decays over training)

- Early: high λ forces the policy to build label coverage.
- Later: low λ lets the policy focus purely on accuracy.



Two Practical RL Variants

Simple tabular version (RDES/B)

- Tabular states, simple and efficient
- Great for intent classification

Neural policy version

- Neural nets for large state spaces
- Stronger on hard reasoning tasks

RDES/C —usually the strongest

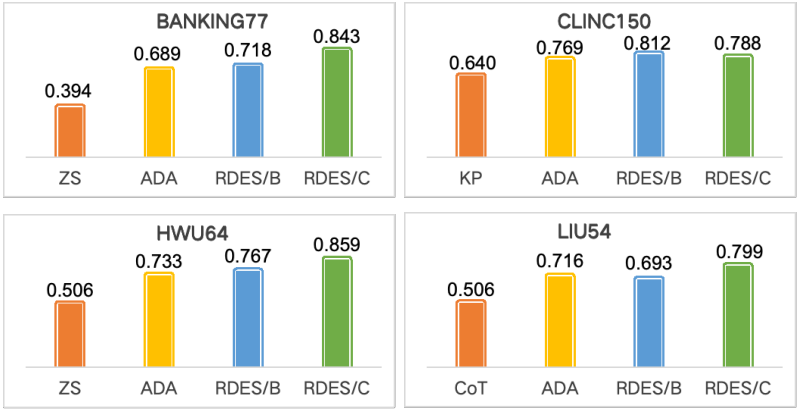
Standard RDES plus Chain-of-Thought reasoning on the selected demonstrations. This variant wins most often across tasks and models.

Experimental Setup

Key facts at a glance:

Axis	**Configuration**
Tasks	4 intent (Banking77, CLINC150, HWU64, Liu54) + reasoning (GSM-8K, SST5, BBH subsets)
LLMs	14 models (GPT-3.5-turbo, Doubao, Hunyuan, Gemma-2, LLaMA-3, Qwen-2.5, DeepSeek-R1, ...)
Baselines	Prompt engineering (ZS, KP, L2M, CoT, Self-Refine) + selection (FS, FSC, AES, RDS, ADA)
Protocol	Default $k = 5$; sensitivity study at $k \in \{3, 5, 7, 10\}$; unified feature pipeline
Metrics	Accuracy, label coverage, k-robustness, runtime overhead

Main Results: Consistent Gains on Intent Classification



Closed + open-source averages (4 datasets). RDES (especially RDES/C) beats all baselines.

Key Quantitative Results

Closed-source

	B77	C150	HWU	LIU
Best base	0.565	0.780	0.734	0.590
RDES/B	0.737	0.835	0.748	0.707
RDES/C	**0.838**	**0.902**	**0.872**	**0.824**

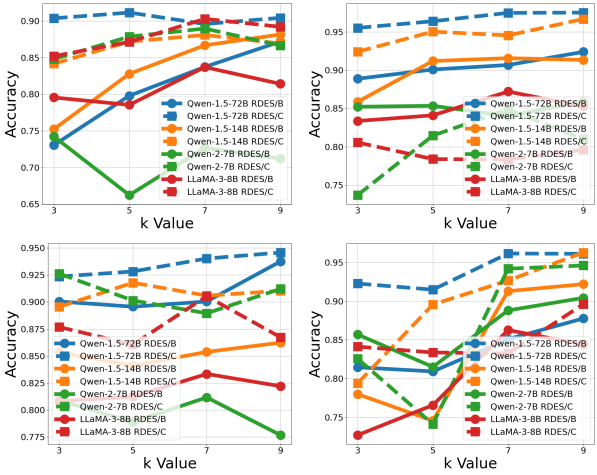
Open-source

	B77	C150	HWU	LIU
Best base	0.752	0.763	0.734	0.779
RDES/B	0.708	**0.800**	0.776	0.686
RDES/C	**0.845**	0.731	**0.853**	**0.786**

Headline Numbers

- Closed-source: **+12.8%** average gain, RDES/C wins **all 4**
- Open-source: RDES/C leads 3/4; consistent gains on 8+ models (RDES/B best on CLINC150)
- Runtime overhead stays below **0.1%**

Robustness: Performance Stays Stable Across Different k



Performance stays stable from $k = 3$ to $k = 10$. RDES works reliably across different demonstration budgets and does not overfit to one particular size.

No brittle k-tuning needed

- Effective even when annotation budget is uncertain.
- No need to search for the "optimal" number of demos.

Extension: Works on Hard Reasoning Too

Tasks & Models

Reasoning tasks: GSM-8K, SST5, Big-Bench Hard (Boolean expressions and Web of Lies).
Tested on strong models: Qwen-2.5-72B and DeepSeek-R1-32B.

Key Takeaway

- RDES/C and the neural variant remain competitive or win.
- The neural version shines when exploration matters most.
- The relevance–diversity RL idea generalizes far beyond intent classification.

Contributions

1. **Formulation** —First to formulate ICL demonstration selection as a query-adaptive sequential decision process whose reward explicitly balances relevance and diversity.
2. **Algorithms** —Two practical RL backbones (tabular and neural) plus easy integration with Chain-of-Thought (RDES/C).
3. **Evidence** —+12.8% average gain on 4 intent tasks across 14 LLMs, plus gains on reasoning tasks, all with $<0.1\%$ overhead.
4. **Insight** —We learn a lightweight *input composition policy* while keeping the LLM frozen —a practical alternative to heavy fine-tuning.

Takeaways

Three Messages

1. Similarity-only selection is fundamentally limited for robust ICL.
2. RDES turns static retrieval into a learnable, query-aware policy that jointly optimizes relevance and diversity via RL.
3. Delivers large gains in accuracy and label coverage, with almost zero overhead. Easy to plug into existing LLM pipelines.

One Sentence

RDES converts static retrieval into adaptive, feedback-driven selection for reliable in-context learning.

Thank you. Questions?