# Instantly Learning Preference Alignment via In-Context DPO

**Song et al., 2025**

**Review by Jason Gillette**


## Introduction

Paper summary

- Proposes *In-Context Direct Preference Optimization (ICDPO)* — a *tuning-free* alignment method.
- Integrates **in-context learning (ICL)** with **Direct Preference Optimization (DPO)**.
- Alignment *without fine-tuning* by conditioning LLMs on retrieved preferences.
- Matches DPO accuracy while cutting compute and storage costs.

Key points

- No parameter updates — alignment entirely from prompt context.
- Uses *expert–amateur collaboration* to simulate fine-tuning.
- Employs a two-stage retriever (BM25 → SBERT).
- Outperforms supervised fine-tuning and rivals DPO.

# Background

## Human Preference Alignment

- Ensures model outputs are *helpful, harmless, honest*.
- Prevents harmful or unsafe completions.
- Traditionally achieved via **RLHF**, which uses:
    1. Base model ($\pi_0$)
    2. Reward model (r)
    3. Policy model ($\pi$) optimized with PPO.

---

## In-Context Learning (ICL)

- Model "learns" behaviors from examples *within the prompt* — no weight updates.
- Behaves as if fine-tuned temporarily:

> *Condition → predict → adapt instantly.*

- Enables alignment through *prompt design* instead of retraining.

---

## Direct Preference Optimization (DPO)

- Simplifies RLHF by removing the reward model.
- Optimizes policy directly from human preference pairs (y+, y-).
- Grounds responses the base model would also deem probable (KL-regularized).
- Loss encourages higher likelihood for preferred over dis-preferred responses.
- Method still depends on additional training.

---

# Contributions

- **ICDPO Framework** — merges DPO's preference logic with ICL.
- **Tuning-free Alignment** — performs DPO entirely in-context.
- **Expert–Amateur Collaboration** — contrastive scoring between conditioned and unconditioned model states.
- **Two-Stage Retrieval System** — BM25 + SBERT for efficient, relevant demonstration selection.
- **Comparable Accuracy, Fractional Cost** — achieves DPO-level results using only inference.

---

# Proposed Architecture

- Retriever R:
    - **BM25** → fast lexical candidate selection.
    - **SBERT** → semantic reranking for relevance.

- Prompt Builder: inject top-k demonstrations (prompt, preferred, dispreferred).
- LLM Scoring:
    - **Expert Score S(x,y)** — log-probability with demonstrations.
    - **Amateur Score $\hat{S}(x,y)$** — log-probability without demonstrations.
    - **$\Delta(x,y)=S-\hat{S}$** → improvement from context.
    - **$D(x)=\Delta(y^{+})-\Delta(y^{-})$** → preference signal.
- Decision: choose or rerank by D(x).

*(See diagram of ICDPO architecture)*

---

# Methods / Experiment Set-up

## Datasets

- **HH-RLHF (Anthropic Helpful–Harmless)** Human-labeled preference pairs for *helpfulness* and *harmlessness*; used for reward model and GPT-4 evaluation.
- **AlpacaEval (Li et al., 2023b)** GPT-4–based benchmark for *instruction following*; 805 test samples and 17 701 human-labeled demonstrations. Reports win/tie/lose rates and length-controlled bias correction.

---

## Models Tested

- **LLaMA-7B**
- **LLaMA-2-7B**
- **Mistral-7B-v0.1**

Chosen to test how ICDPO generalizes across architectures of varying strength.

---

## Baselines

Tuning-free / Prompt-based methods

- *Zero-Shot* – no alignment
- *RM-BoN* – Best-of-N via reward-model scoring
- *RM-Aug* – reward-guided decoding
- *URIAL* – retrieved in-context alignment
- *RAIN* – self-evaluative ICL with feedback loop

Fine-tuning methods

- *SFT* – supervised on preferred answers
- *DPO* – direct preference optimization (trained)
- *ICDPO* – in-context DPO (no training, retrieval + scoring)

---

### Evaluation Metrics

- **RMtest score** – alignment strength measured by a trained reward model.
- **GPT-4 Win/Tie/Lose rate** – GPT-4 acts as judge vs. reference responses.
- **Length-Controlled Win Rate** – bias-corrected GPT-4 metric (AlpacaEval).
- **Mean Reciprocal Rank (MRR)** – agreement between ICDPO and GPT-4 rankings.

---

## Results — RM Evaluations

| Method | LLaMA | LLaMA-2 | Mistral |
|---|---|---|---|
| Zero-Shot | −36.5 | −30.7 | −11.8 |
| RM-BoN | −31.0 | −22.8 | 0.5 |
| RM-Aug | −27.7 | −24.6 | 3.3 |
| **ICDPO** | 25.6 | 62.3 | 68.8 |
| **ICDPO + Ŝ** | 28.5 | 63.7 | 71.2 |
| **ICDPO + ŜR (two-stage retriever)** | **51.6** | **69.7** | **73.6** |

*ICDPO variants outperform reward-based and prompt baselines, matching or exceeding DPO alignment.*

---

## Results — GPT-4 and AlpacaEval

GPT-4 Evaluation (200 HH-RLHF samples)

- ICDPO > RM-Aug > URIAL > Zero-Shot.
- Reliable correlation with RMtest scores.
- Helpful tasks harder than Harmless.

AlpacaEval (Instruction Following)

| Base | RM-BoN | RM-Aug | URIAL | RAIN | ICDPO | ICDPO + Ŝ |
|---|---|---|---|---|---|---|
| LLaMA | 1.6 | 2.3 | 5.4 | 6.8 | **10.0** | **10.3** |
| LLaMA-2 | 6.3 | 6.2 | 7.0 | 16.3 | **18.7** | **19.2** |
| Mistral | 17.1 | 18.5 | 21.9 | 26.3 | **26.5** | **28.3** |

*ICDPO + Ŝ achieves top length-controlled win rates on all models.*

---

## Ablation Study Highlights

- Removing contrastive score Ŝ ⇒ large alignment drop.
- Removing SBERT reranker ⇒ poor retrieval quality.
- Using random demos ⇒ performance near SFT.
- Increasing demos beyond ≈ 5 ⇒ no gains (saturation).

- High-quality retrieval is essential for ICDPO success.

---

## Potential Weaknesses

- Reliance on retriever quality — poor matches harm alignment.
- Unclear robustness to domain shift or safety edge cases.
- No true weight updates → benefits are temporary.
- Still requires annotated human preference data.
- Limited evaluation scope (only two datasets and two models).

---

## References

- Instantly Learning Preference Alignment via In-context DPO (Song et al., NAACL 2025)