ICLR 2026 Oral

Semi-Supervised Preference Optimization with Limited Feedback

SSPO learns from a small set of paired preference labels together with a large pool of unpaired responses deriving a principled reward threshold for pseudo-labeling unpaired data, enabling strong alignment with much less human feedback.

Seonggyun Lee[1], Sungjun Lim[1], Seojin Park[2], Soeun Cheon[2], Kyungwoo Song[1]*

[1] Yonsei University [2] KAIST * Corresponding Author.

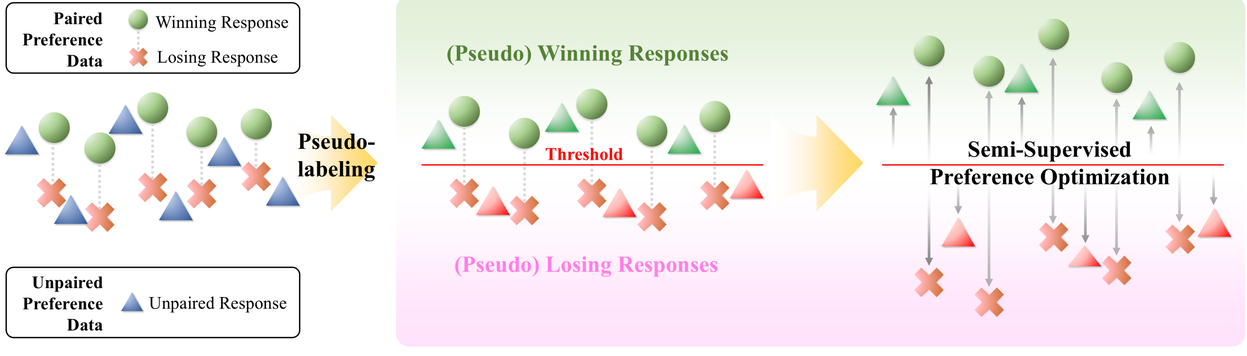
The Alignment Bottleneck: High Cost of Paired Preference Data

5-10 min.
average **time** of annotation **per data point** by human

$10-30
average **cost** of annotation **per data point** by human

- **Prohibitive Costs:** Current methods rely heavily on paired feedback, which requires expert labor averaging 5-10 minutes per comparison and costing $10-30 per data point.
- **Risks of Synthetic Data:** While LLM-based self-annotation is an alternative, it risks creating feedback loops that propagate model biases and lack context-dependent nuances.
- **Underutilized Assets:** Existing domain-specific SFT datasets are rich in expert knowledge but lack the explicit preference labels required for standard PO.

Our Solution: Semi-Supervised Preference Optimization (SSPO)

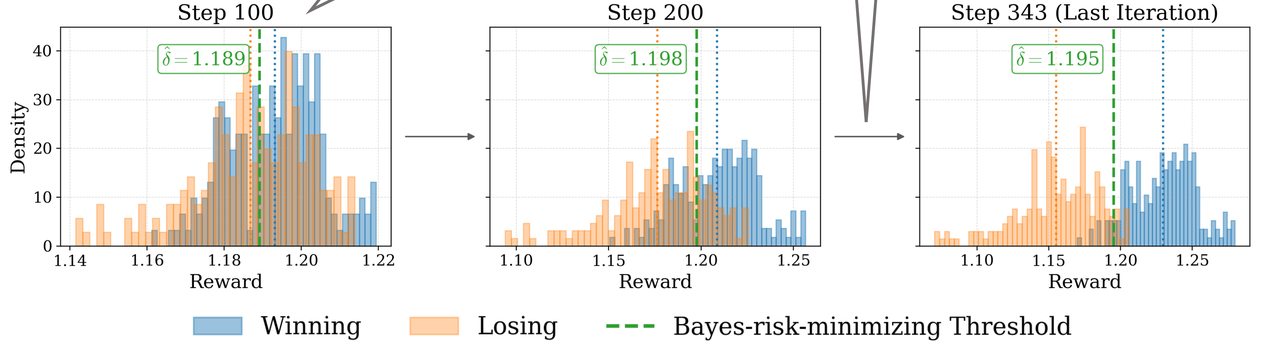


- **Data Efficiency:** SSPO distills latent preferences from large-scale unpaired data, drastically reducing acquisition costs while maintaining human alignment.
- **Bayes Risk Minimization:** We numerically solve for the threshold that minimizes the estimated Bayes risk to assign pseudo-labels to unpaired responses, with a threshold-based boundary (Theorem 1).

How does SSPO work?

High overlap between winning and losing rewards. The model prioritizes paired data, providing a stable anchor to prevent noise amplification.

As training progresses, the model prioritizes unpaired data, allowing the dynamic threshold to accurately pseudo-label with high confidence.



SSPO trained on just 1% of paired data surpasses strong baselines trained on 10%.

Trained on **UltraFeedback**	Data Size	Phi-2 (2.7B)		Mistral-7B-IT		Llama3-8B-IT	
		LC	MT	LC	MT	LC	MT
DPO	1%	3.6	**6.3**	17.0	7.6	12.1	**8.0**
	10%	4.6	**6.3**	18.0	7.6	11.0	**8.0**
SSPO	1%	**7.2**	**6.3**	**26.7**	**7.7**	**15.0**	**8.0**
	10%	**7.7**	**6.3**	**30.0**	**7.7**	**20.7**	7.9

LC : Length-Controlled Win Rate
MT : MT-Bench

Paper (OpenReview)