EPM-RL: Reinforcement Learning for On-Premise Product Mapping in E-Commerce

Minhyeong Yu[1], **Seunghyun Lee**[1], **Wonduk Seo**[1,2*†]

[1]Enhans, Seoul, South Korea
[2]Peking University, Beijing, China
minhyeong@enhans.ai, seunghyun@enhans.ai, seowonduk@pku.edu.cn

Abstract

Product mapping decides whether two e-commerce listings refer to the same product. We propose **EPM-RL**, an on-premise framework that distills LLM comparison reasoning into a local model via PEFT and RL, improving balanced matching under privacy and cost constraints. On an internal benchmark, **EPM-RL** yields the strongest F1 by boosting recall while maintaining competitive precision.

Motivation

Product mapping is difficult because listing titles are noisy and can be lexically similar even for different SKUs (Aanen, Vandic, and Frasincar 2015). Lightweight rule-based and encoder-based systems are efficient, but they often mishandle these hard cases. LLM reasoning, including retrieval-augmented generation (RAG), helps resolve variant-level conflicts but is costly and privacy-sensitive (Seo et al. 2025). We distill this reasoning into a single on-premise model.

Framework

EPM-RL trains a local matcher in two stages. (1) We distill short comparison rationales from a strong teacher and fine-tune a student with LoRA (Hu et al. 2022) to generate a rationale plus a binary match label. Given input x_i and target output o_i, the supervised objective is

$$\mathcal{L}_{\mathrm{SFT}} = -\sum_{i=1}^{N}\sum_{t=1}^{T_i} \log P_\theta(o_{i,t} \mid x_i, o_{i,<t}).$$

(2) We further optimize the student with Group Relative Policy Optimization (Liu et al. 2024), using rewards for format compliance, label correctness, and judge-scored reasoning quality. The RL objective is

$$\mathcal{L}_{\mathrm{RL}} = -\mathbb{E}_x \left[\frac{1}{K} \sum_{k=1}^{K} \hat{A}^{(k)} \log \pi_\theta(o^{(k)} \mid x) \right].$$

The judge aggregates scores for core identity, identifier consistency, and variant conflicts:

$$R(o,x) = \frac{s_f + 2s_y + \frac{s_{\mathrm{core}}+s_{\mathrm{id}}+s_{\mathrm{var}}}{3}}{4}.$$

*Wonduk Seo is the corresponding author for this work.

†This work was done while Wonduk was working at Enhans.

Method	Acc.	Prec.	Rec.	F1
RoBERTa-base	0.837	0.741	0.733	0.737
Zero-shot Classification	0.854	0.811	0.657	0.726
Chain-of-Thought (CoT)	0.845	0.864	0.561	0.680
Entity-Attr.	0.844	0.875	0.548	0.674
Multi-Agent RAG	**0.864**	0.891	0.612	0.726
GPT-5.4 Reasoning	0.863	**0.976**	0.548	0.702
LoRA + Reasoning (Ours)	0.857	0.710	0.867	0.781
EPM-RL (Ours)	0.845	0.782	**0.874**	**0.812**

Table 1. Main results on the test set. LLM baselines are compared using the same model (30B Scale).

Experiments

We evaluate **EPM-RL** on an internal benchmark of ∼12K labeled product pairs across ∼500 brands. We compare against encoder baselines and prompt-only or retrieval-augmented LLM setups under a fixed 30B-scale setting. Table 1 shows that **EPM-RL** achieves the best F1 by improving recall while maintaining competitive precision.

Conclusion

EPM-RL converts expensive agentic product-mapping reasoning into a scalable on-premise model. By combining structured reasoning distillation, LoRA fine-tuning, GRPO, and judge-based rewards, it provides an efficient and inspectable alternative to inference-time multi-agent systems.[1]

References

Aanen, S. S.; Vandic, D.; and Frasincar, F. 2015. Automated product taxonomy mapping in an e-commerce environment. *Expert Systems with Applications*, 42(3): 1298–1313.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2): 3.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Seo, W.; Shin, T.; An, H.; Kim, D.; and Lee, S. 2025. Question-to-Knowledge (Q2K): Multi-Agent Generation of Inspectable Facts for Product Mapping. In *2025 IEEE International Conference on Big Data (BigData)*, 2646–2653.

[1]Detailed technical report: https://arxiv.org/abs/2604.23993.