Can Compact Language Models Search Like Agents? Distillation-Guided Policy Optimization for Preserving Agentic RAG Capabilities

Rikuto Kotoge[*1] Mai Nishimura[2] Jiaxin Ma[2]
[1]The University of Osaka [2]OMRON SINIC X Corporation *work done as an intern at OMRON SINIC X.

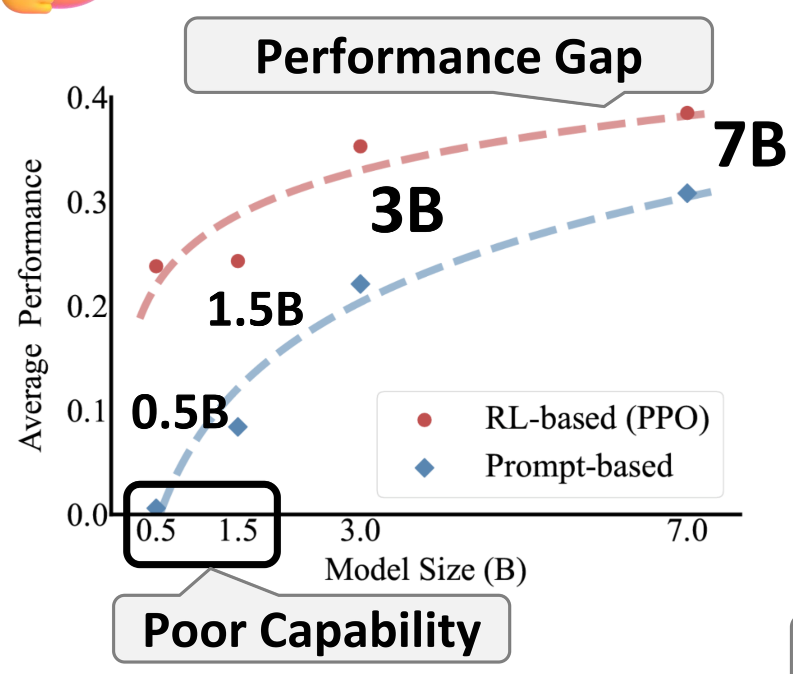
ACL 2026 SAN DIEGO JULY 2–7

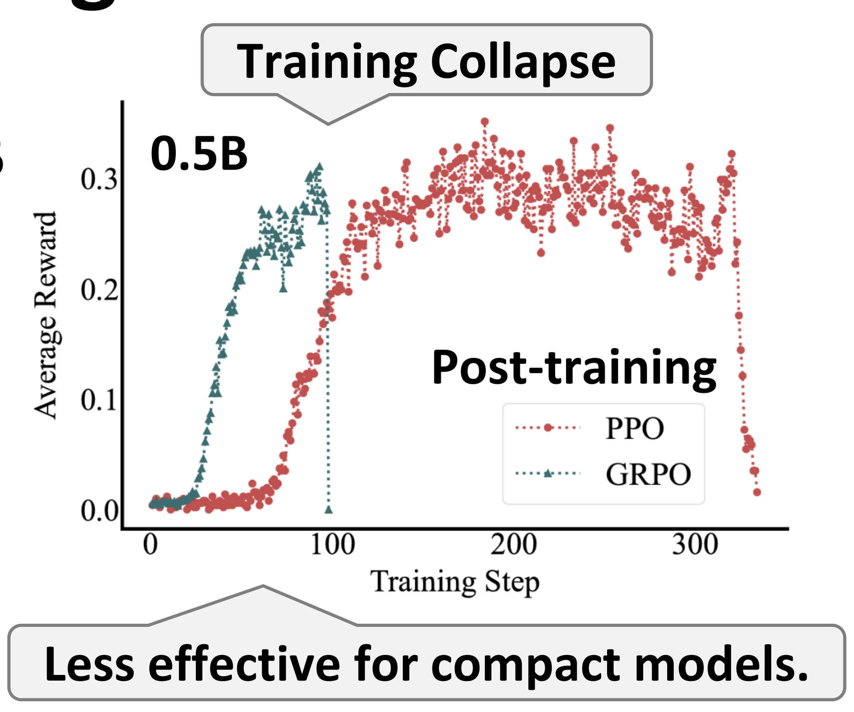
Motivation

- Agentic RAG requires reasoning, search, and answer synthesis.
- Existing methods rely on large models, making agentic RAG inaccessible on computing resource-limited environments.

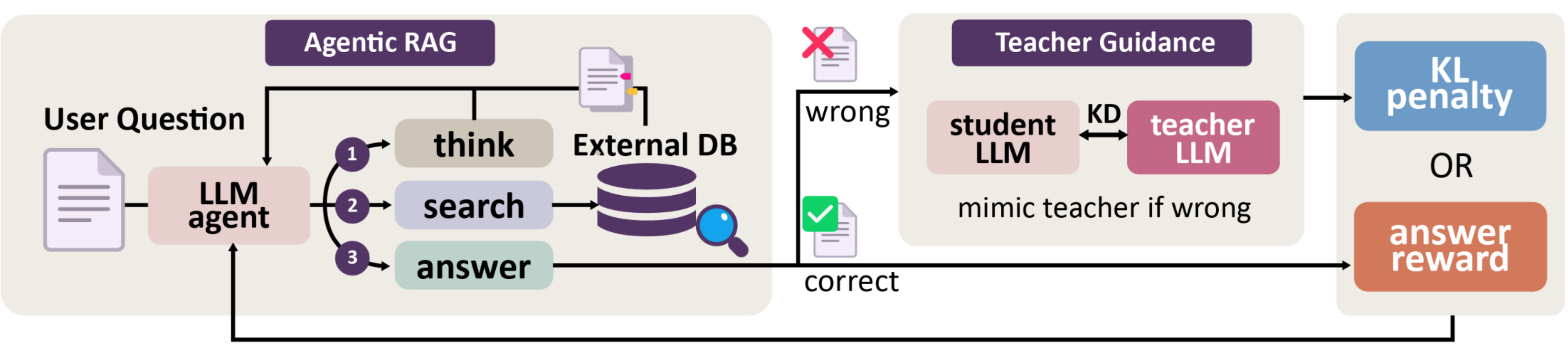
Goal Enabling compact LLMs (0.5-1B) to acquire agent-like behaviors

Technical challenges

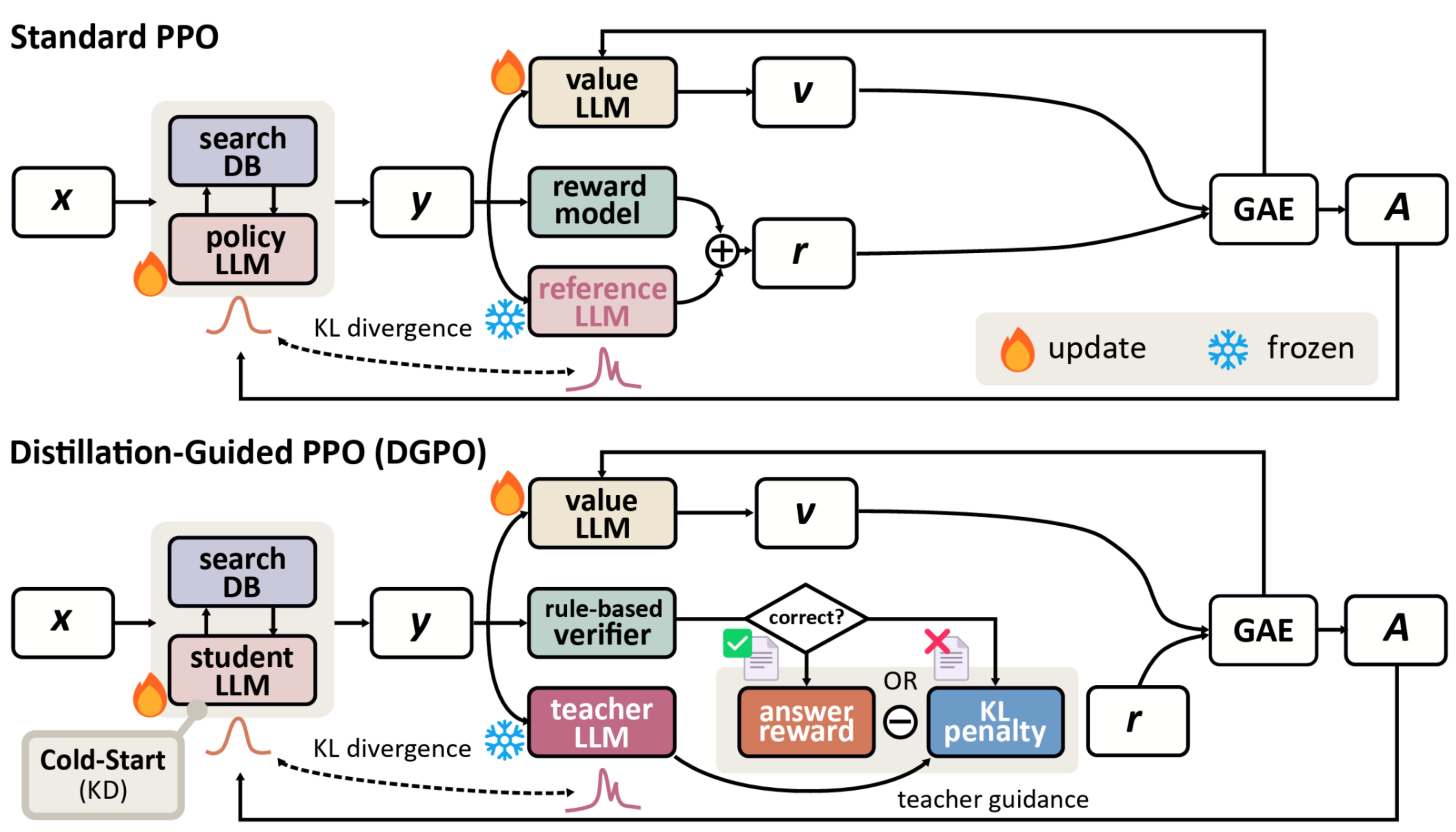




Distillation-Guided Policy Optimization (DGPO) Proposed Method

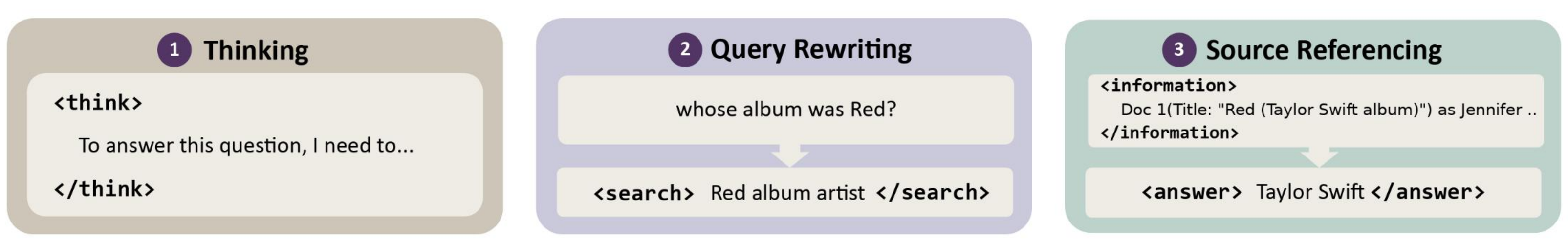


Key Idea "Targeted correction on incorrect attempts, self-exploration on correct attempts."



1. **Cold-start Distillation (KD)**
   - Stabilizes early learning with **Knowledge Distillation (KD)**.
2. **PPO with Selective Teacher Guidance**
   - Reward if right, mimic teacher if wrong.

Agentic RAG Capability (ARCap) Novel Metric



We propose ARC to evaluate three core process capabilities:
(1) Thinking: Multi-hop reasoning, when-to-search decisions.
(2) Query Rewriting: Reformulate questions into search queries.
(3) Source Referencing: Extract and incorporate retrieved information.

Experiments

Q1. Do our compact models preserve the overall performance of the teacher models?

Qwen 2.5 (3B → 0.5B)	NQ	TriviaQA	PopQA	HotpotQA	2wiki	MuSiQue	Bamboogle	Avg.
Student-0.5B	0.004	0.006	0.007	0.007	0.015	0.000	0.000	0.006
Teacher-3B	0.365	0.569	0.393	0.340	0.368	0.135	0.298	0.353
PPO (Jin et al., 2025)	0.306	0.444	0.379	0.205	0.218	0.041	0.073	0.238
SFT → PPO	0.338	0.415	0.359	0.296	0.275	0.088	0.250	0.289
GKD (Agarwal et al., 2024)	0.266	0.408	0.358	0.216	0.217	0.055	0.161	0.240
SeqKD (Kim and Rush, 2016)	0.331	0.416	0.364	0.283	0.273	0.089	0.169	0.275
KD (Hinton et al., 2015)	0.331	0.431	0.373	0.286	0.284	0.091	0.290	0.298
DistiLLM (Ko et al., 2024)	0.333	0.442	0.373	0.288	0.270	0.095	0.209	0.287
TAID (Shing et al., 2025)	0.325	0.427	0.365	0.290	0.270	0.079	0.218	0.282
DGPO (ours)	**0.378**	0.481	**0.402**	**0.342**	0.303	0.120	0.274	0.329

- ✓ DGPO achieves **55× improvement over base model** (0.006 → 0.329).
- ✓ Approaches teacher performance (0.353).
- ✓ **Surpasses the teacher** on 3 datasets (NQ, TriviaQA, HotpotQA).

Q2. How well do compact models retain individual ARCap

Models	NQ		MuSiQue	
Qwen2.5 (3B → 0.5B)	w/o	w/ thinking	w/o	w/ thinking
Student-0.5B	0.386	0.034	0.166	0.013
Teacher-3B	0.589	0.560	0.413	0.357
PPO	0.547	0.581	0.258	0.242
KD	0.540	0.544	0.321	0.256
DGPO	0.565	0.593	0.312	0.287

Table 4: Source referencing and thinking performances on NQ and MuSiQue datasets.

Models	NQ (first hop)	MuSiQue (multi-hop)	
Qwen2.5 (3B → 0.5B)	Hit ratio	Hit ratio	Search steps
Student-0.5B	0.004	0.052	3.86
Teacher-3B	0.682	0.668	1.60
PPO	0.711	0.568	1.68
KD	0.675	0.570	2.45
DGPO	0.682	0.583	2.64

Table 5: Query rewriting performance on NQ and thinking performance on MuSiQue datasets.

(a) Source Referencing: DGPO provides **strong information extraction** when the correct evidence is directly available.

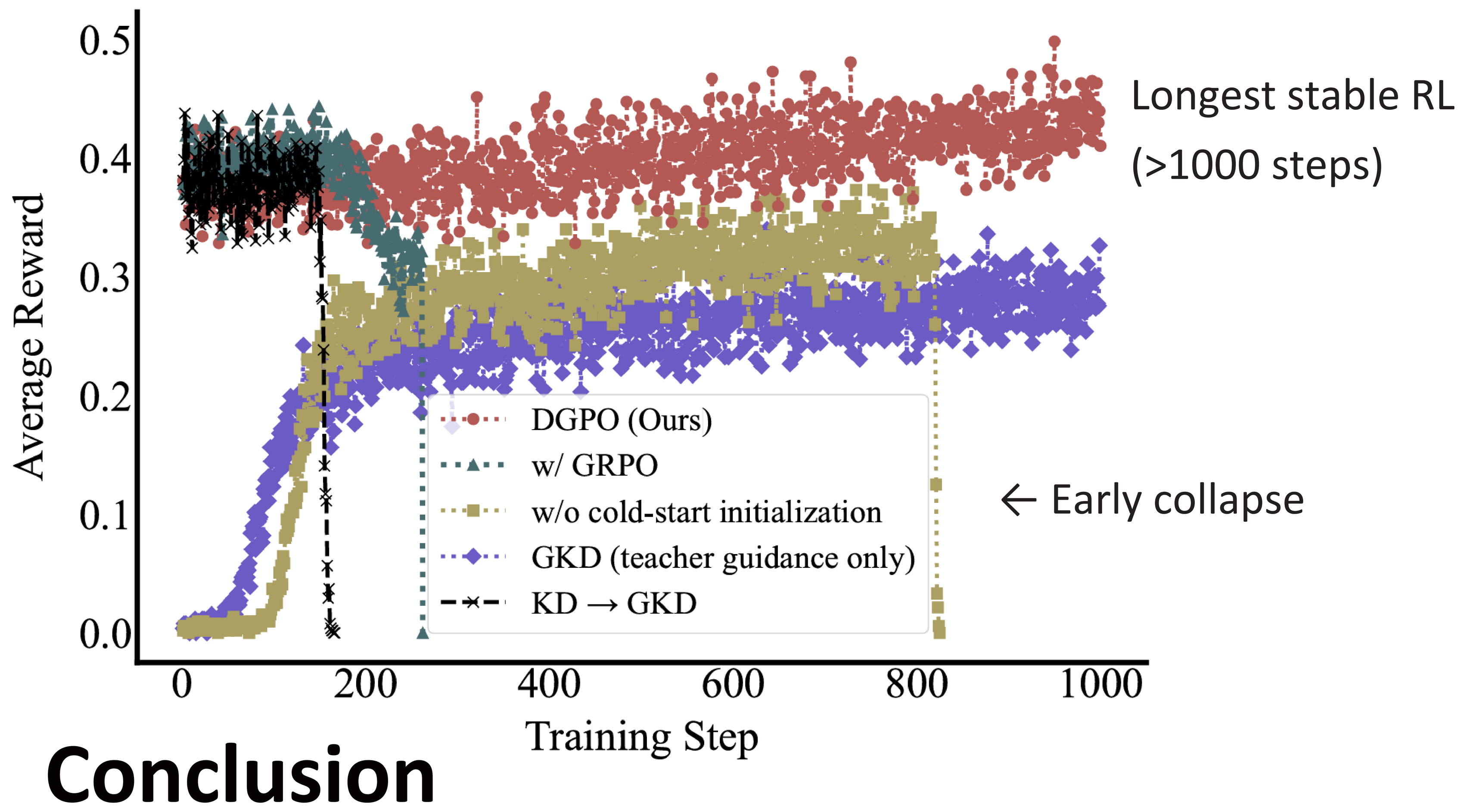
(b) Query Rewriting: DGPO achieves **teacher-level** query rewriting.

(c) Thinking: DGPO exhibits the **strongest multi-hop reasoning** by **taking more search steps** than the teacher model.

Q3. Which components of our method contribute most to performance improvements?

Method	init (KD)	pipeline	KL penalty	NQ	TriviaQA	PopQA	HotpotQA	2wiki	MuSiQue	Bamboogle	Avg.
DGPO	✓	KD → PPO	selective	0.378	0.481	0.402	0.342	0.303	0.120	0.274	0.329
(a) w/o cold-start initialization	–	KD → PPO	selective	0.370	0.465	0.394	0.330	0.299	0.117	0.266	0.320
(b) w/o selective kl penalty	✓	KD → PPO	uniform	0.362	0.464	0.394	0.323	0.306	0.114	0.234	0.314
(c) w/o teacher guidance	✓	KD → PPO	–	0.353	0.455	0.384	0.316	0.287	0.098	0.250	0.306
(d) invert pipeline order	–	PPO → KD	–	0.320	0.426	0.371	0.287	0.282	0.084	0.234	0.286

Q4. Does our method mitigate training instability and avoid performance plateau in compact models?



Conclusion

- ✓ DGPO unlocks agentic RAG for compact LLMs.
- ✓ Compact models can **exceed teacher models** on several benchmarks.
- ✓ Enables on-device search agents and broad accessibility.
- ✓ Provides a **foundation for future agentic reasoning** in compact models.