

RL without Tears:

Appendix: Datasets and Systems

Chenglong Wang

Northeastern University, China

WANGCHENGLONG@MAIL.NEU.EDU.CN

Hang Zhou

Northeastern University, China

STCEUM@GMAIL.COM

Tongran Liu

Institute of Psychology, Chinese Academy of Sciences, China

LIUTR@PSYCH.AC.CN

Jingbo Zhu

Northeastern University, China

ZHUJINGBO@MAIL.NEU.EDU.CN

Tong Xiao

Northeastern University, China

XIAOTONG@MAIL.NEU.EDU.CN

Appendix: Datasets and Systems

Dataset Name	Use Case			Scale	Modality	Feedback / Reward Signal	
	★ RM	🏠 Rsn.	👤 Agent			Source	Category
Anthropic/hh-rlhf	✓	✗	✗	169K	📄	Human	Pairwise preference
stanfordnlp/SHP-2	✓	✗	✗	4.8M	📄	Human / crowd	Pairwise preference
H4/stack-exchange-preferences	✓	✗	✗	10.7M	📄	Crowd score	Score / ranking
Summarize from Feedback	✓	✗	✗	64.8K	📄	Human	Pairwise preference
UltraFeedback	✓	✗	✗	64K / 340K	📄	GPT-4	Scores, critiques, pairs
berkeley-nest/Nectar	✓	✗	✗	183K / 3.8M	📄	GPT-4	7-way ranking
Skywork-Reward-Preference-80K	✓	✗	✗	80K	📄	Mixed / curated	Pairwise preference
nvidia/HelpSteer	✓	✗	✗	37K	📄	Human	Attribute scores
nvidia/HelpSteer2	✓	✗	✗	10K pairs	📄	Human	Attribute scores / preference
nvidia/HelpSteer3	✓	✗	✗	40K pref.	📄	Human	Preference, feedback, edits
Arena Human Preference 55K	✓	✗	✗	55K	📄	Human users	Arena battle preference
VLFeedback	✓	✗	✗	80K / 380K	📄📺	GPT-4V	Multimodal preference
RLHF-V-Dataset	✓	✗	✗	5.7K	📄📺	Human	Fine-grained correction
openbmb/RLAIF-V-Dataset	✓	✗	✗	83K	📄📺	AI feedback	Pairwise preference
OpenGVLab/MMPR-v1.2	✓	✓	✗	3M	📄📺	AI / verifier	Multimodal reasoning preference
open-r1/OpenR1-Math-220k	✗	✓	✗	220K	📄	Verifier / judge	Math reasoning traces
AI-MO/NuminaMath	✗	✓	✗	900K	📄	Rule / GPT-4	Math CoT / TIR
PRIME-RL/Eurus-2-RL-Data	✗	✓	✗	482K	📄	Verifier	Math/code outcome reward
AgentGym-RL-Data	✗	✓	✓	184K	📄	Env.	Multi-turn agent reward
ALFWorld	✗	✗	✓	3.5K train	📄	Env.	Text-world task reward
WebShop	✗	✗	✓	12K tasks	📄	Env.	Web shopping reward
Search-R1 Data	✗	✓	✓	170K train	📄	Retriever / rule	Search-tool QA reward
Sokoban	✗	✓	✓	Env. gen.	📄📺	Env.	Puzzle-solving reward
Gym Cards	✗	✓	✓	Env. gen.	📄📺	Env.	Logic-game reward
ToolBench	✗	✓	✓	126K	📄	API execution	Tool-use trajectories

Table 1: Preference, reasoning, and trainable agentic RL datasets and environments for LLMs. Here, RM denotes reward modeling, Rsn. denotes reasoning, Env. denotes environment-based feedback, Env. gen. denotes procedurally generated environments, and TIR denotes tool-integrated reasoning.

System Name	Supported Modality				Supported Training Approaches
	Text	Image	Video	Audio	
TRL	✓	✓	✗	✗	SFT, Reward Modeling, PPO, GRPO, RLOO, DPO, KTO, ORPO, Online-DPO; Agentic RL via OpenEnv/Harbor.
OpenRLHF	✓	✓	✗	✗	SFT, RM, PPO, REINFORCE++, GRPO, RLOO, Dr.GRPO, DPO, KTO; Agentic RL with single-turn/multi-turn executors.
verl	✓	✓	✗	✗	SFT, PPO, GRPO, GSPO, ReMax, REINFORCE++, RLOO, DAPO, Dr.GRPO; Agentic RL via multi-turn tool calling.
verl-agent	✓	✓	✗	✗	GiGPO, GRPO, PPO, DAPO, GSPO, RLOO, REINFORCE++, LoRA; purpose-built Agentic RL for long-horizon LLM/VLM agents.
EasyR1	✓	✓	✗	✗	GRPO, DAPO, REINFORCE++, ReMax, RLOO, GSPO, CISPO; no native Agentic RL.
LLaMA-Factory	✓	✓	✓	✓	Pre-training, SFT, RM, PPO, DPO, KTO, ORPO, SimPO; no native Agentic RL.
VeOmni	✓	✓	✓	✓	Single-/multi-modal pre-training and post-training, SFT-style training, DPO, RL trainer backend; no native Agentic RL environment stack.
ms-swift	✓	✓	✓	✓	Pre-training, SFT, RM, PPO, GRPO, DPO, KTO, ORPO, CPO, SimPO, GKD; Agentic RL via multi-turn GRPO/tool-use training.
Align-Anything	✓	✓	✓	✓	SFT, RM, PPO, DPO, KTO, ORPO, SimPO, rule-based RL; Agent RL on roadmap.
OpenRLHF-M	✓	✓	✗	✗	PPO, GRPO, RLOO, Online-RLHF, Rejection Sampling for multimodal models; no native Agentic RL.
DeepSpeed-Chat	✓	✗	✗	✗	SFT, Reward Modeling, PPO-based RLHF; no native Agentic RL.
ROLL	✓	✓	✗	✗	SFT, DPO, distillation, PPO, GRPO, REINFORCE++, GSPO, RAFT++, StarPO, GiGPO; Agentic RL.
slime	✓	✗	✗	✗	PPO, GRPO, GSPO, REINFORCE++, OPD; Agentic RL via custom generation, tools, sandboxes, and verifier rewards.
OpenClaw-RL	✓	✓	✗	✗	Binary RL/GRPO, OPD, Hybrid RL, LoRA training; asynchronous Agentic RL for personalized agents, terminal, GUI, SWE, and tool-call settings.
SkyRL	✓	✗	✗	✗	GRPO, DAPO, async RL, Tinker-compatible training; Agentic RL for tool-use, search, SQL, and long-horizon tasks.
Agent Lightning	✓	✗	✗	✗	RL, APO, SFT-style optimization hooks; purpose-built Agentic RL for existing agent frameworks.
RAGEN	✓	✗	✗	✗	PPO/StarPO-style multi-turn optimization, rollout filtering, environment rewards; Agentic RL.
Search-R1	✓	✗	✗	✗	PPO, GRPO, REINFORCE for reasoning-search interleaved models; Agentic RL for search/tool use.
ARaL	✓	✗	✗	✗	PPO/GRPO-style asynchronous RL with agent-framework integration; Agentic RL.

Table 2: RL systems for LLM-based and multimodal post-training, including their supported modalities and corresponding training approaches.

References