How the engine makes evals and training data

Combinatorial coverage of situations. A sandbox world with failure modes. A judge validated before training.

01 Axes
Declare what varies: tool, policy clause, world state, fault, persona, history.

02 Cover
Plan cells so every pair of axis values co-occurs at least once [6].

03 Search
Five writers fill the grid. Budget follows new behavior found [7].

04 World
Tools answer from schema-shaped state. Same seed, same reply. Bad calls fail.

05 Rollout
Run the agent on N tasks × n phrasings × k samples. Keep token logprobs.

06 Grade
Conduct rules first, then your judge. Judge scored against gold labels.

07 Cut
Export SFT rows, DPO pairs, GRPO groups, or a reward-model set.

08 Delta
Re-run held-out tasks after training. Paired diff with bootstrap CI.

GENERATION

What is ours

COVERING ARRAY
Every pair of axis values appears together at least once. Most failures are two-factor interactions [6].

ARM SEARCH
Five situation writers, weighted by yield of new behavior signatures. Novelty search, not importance sampling [7].

SANDBOX WORLD
Built from the tool JSON schemas. Deterministic per seed. Unknown id → not found. Schema-echo argument → refused.

CUTS
SFT: reward=1 rows, loss mask on agent turns. DPO: pairs with margin. GRPO: mixed groups, 20–80% band [4, 5]. RM: all graded rows.

MEASUREMENT

What we take from the literature

pass@1 / pass^k / pass@k
Headline, reliability, RL headroom. Unbiased combinatorial estimators over k samples per task [2, 3].

INTERVALS
Bootstrap over tasks, not rollouts. Before/after: paired difference, permutation p [1]. Ship when the CI excludes 0.

JUDGE
Agreement and κ against gold labels. Length perturbation. Different model family than the policy [8].

HACK SCAN
Var(r) = E[Var(r|task)] + Var(E[r|task]). Only the first term is GRPO gradient. Top within-task feature vs a permutation floor [9].

Importance sampling?
No. IS corrects an estimator for a wrong proposal; we are not estimating production. Each row keeps per-token logprobs, policy version, and temperature, so an async trainer can form the truncated ratio itself [10, 11].

SFT or RL?
Both, from the same graded rows. SFT takes reward=1 rows with a loss mask on agent turns; DPO takes pairs; GRPO takes groups.

Judge is an LLM?
Yes. Measured against gold labels, probed with known hacks, versioned, and a different family than the policy [8].

REFERENCES

[1] Lambert, N. *Reinforcement Learning from Human Feedback*. 2025. rlhfbook.com.
[2] Chen, M. et al. Evaluating Large Language Models Trained on Code. *arXiv:2107.03374*, 2021.
[3] Yao, S. et al. τ-bench. *arXiv:2406.12045*, 2024.
[4] Shao, Z. et al. DeepSeekMath. *arXiv:2402.03300*, 2024.
[5] Yu, Q. et al. DAPO: An Open-Source LLM RL System at Scale. *arXiv:2503.14476*, 2025.
[6] Kuhn, D. R., Wallace, D. R., Gallo, A. M. Software Fault Interactions and Implications for Software Testing. *IEEE TSE* 30(6), 2004.
[7] Lehman, J., Stanley, K. O. Abandoning Objectives: Evolution Through the Search for Novelty Alone. *Evolutionary Computation* 19(2), 2011.
[8] Zheng, L. et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *NeurIPS*, 2023.
[9] Gao, L., Schulman, J., Hilton, J. Scaling Laws for Reward Model Overoptimization. *ICML*, 2023.
[10] Schulman, J. et al. Proximal Policy Optimization Algorithms. *arXiv:1707.06347*, 2017.
[11] Noukhovitch, M. et al. Asynchronous RLHF. *ICLR*, 2025.