Dreamforce

Prospective Hindsight

Self-Calibrating RL via Prediction–Reality Gaps

Jiaxin Zhang, Xiangyu Peng, Qinglin Chen, Yu Li,Hiroaki Hayashi, Chien-Sheng Wu

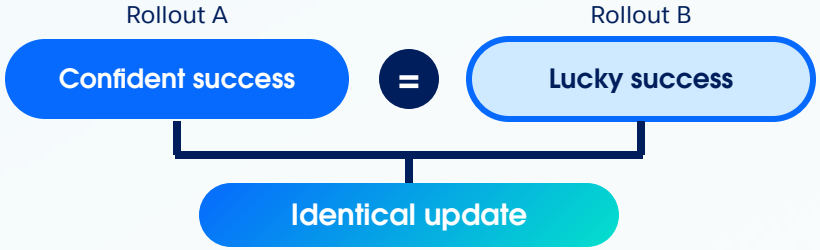
Salesforce AI Research · arXiv:2610.02740



The Hindsight Trap

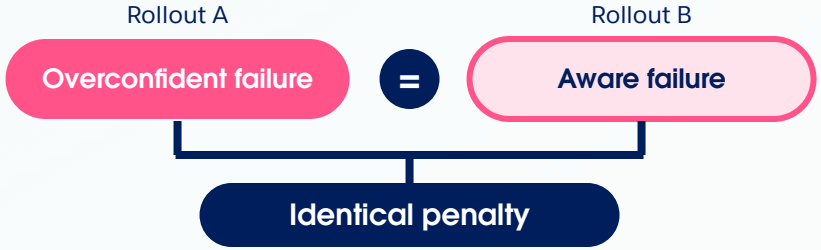
GRPO, OPD, and their combination learn only from what happens **after** acting.

Same success, same reward



Confident success = lucky success. The update cannot tell conviction from chance.

Same failure, same penalty



Overconfident failure = aware failure. The more dangerous mode, and likely to recur.

Gradient is invariant to the prospective belief **z**

$$E(\nabla \ell_{base} \mid z, y) = E(\nabla \ell_{base} \mid y)$$

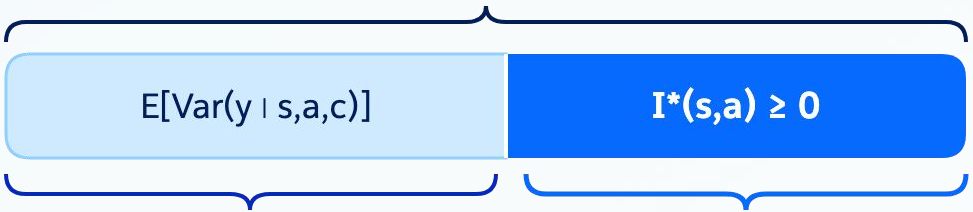
Consequence: retrospective training improves accuracy but does not resolve overconfidence.

The Gap Is Signal, Not Noise

Proposition 1 · Law of total variance

$$\mathrm{Var}(y \mid s,a) = \mathbb{E}[\mathrm{Var}(y \mid s,a,c)] + I^*(s,a)$$

Var(y | s,a) – total outcome variance



aleatoric noise
irreducible even given c

privileged-information gap I*
unresolved without c

I* = outcome variance no action-time predictor can resolve

Rollouts that expose I*

- overconfident failures
- underconfident successes

Agent RL as LUPI

Learning Under Privileged Information (LUPI, Vapnik)

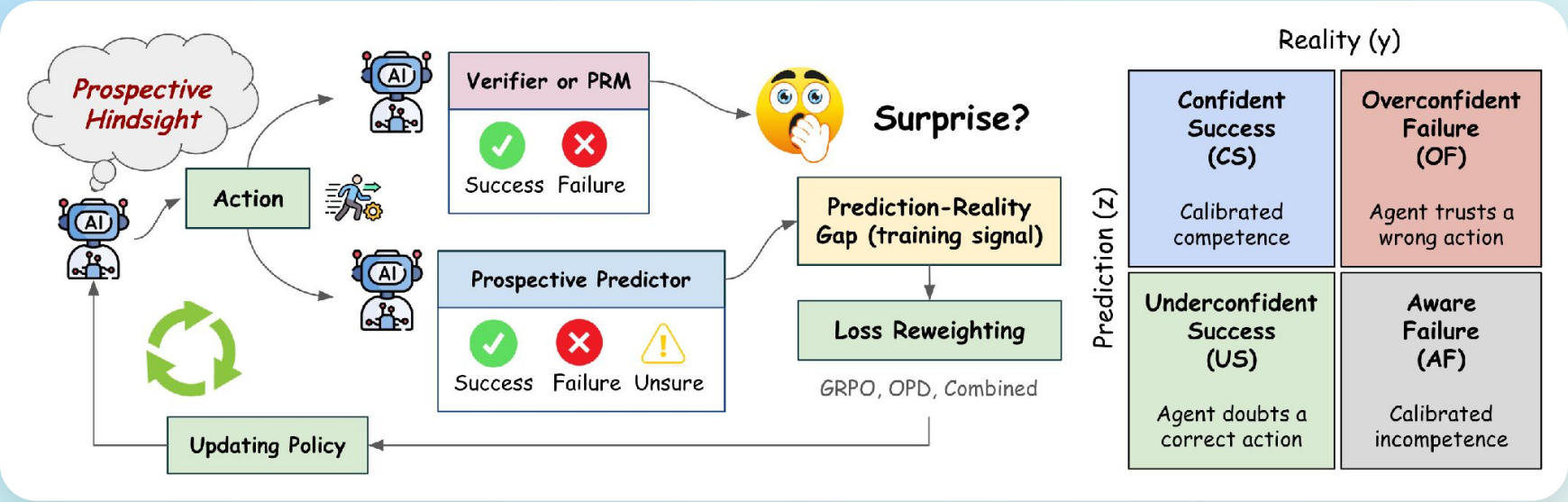
- **(s, a)** standard input
- **y** target
- **c** **privileged info** ground-truth reference or next state; seen by verifier / PRM, not by the agent at action time

TAKEAWAY
Use the belief–outcome disagreement as a training signal

Contributions

01	**Diagnosis**	Identify the **hindsight trap**; link the prediction–reality gap to a **LUPI** (Learning Under Privileged Information) privileged-information quantity.
02	**Method**	**Prospective Hindsight (PH)** turns the gap into a self-calibrating signal; plugs into GRPO, OPD (on-policy distillation), and their combination.
03	**Taxonomy**	Four **rollout cells** (belief × outcome) that decompose calibration during training.
04	**Evidence**	Better task performance **and** calibration, single-turn and multi-turn, across model scales; handles both dominant **miscalibration modes**.

PH at a Glance



1 Act & verify
Verifier / PRM scores the outcome (retrospective)

2 Predict first
Same policy, new prompt, before feedback: Success / Failure / Unsure

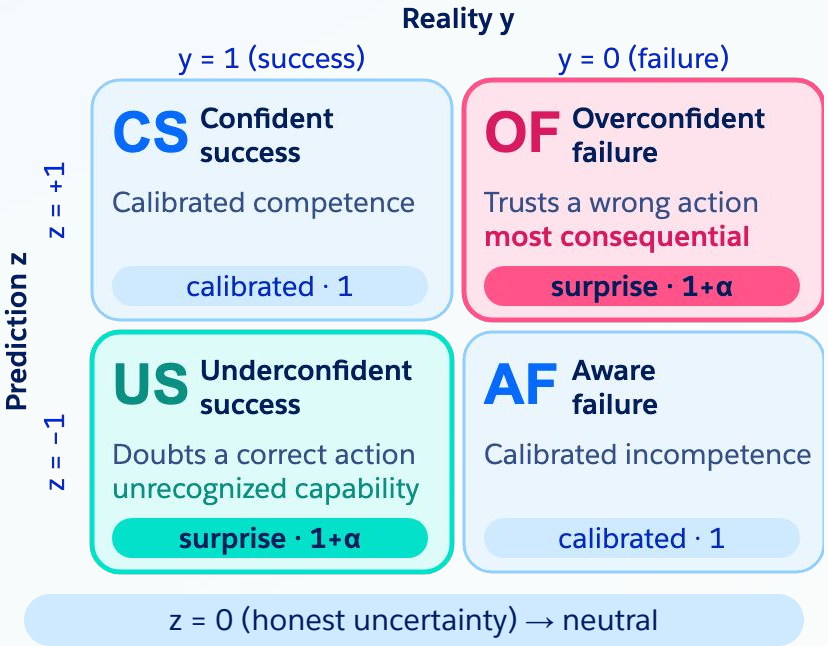
3 Measure surprise
Prediction–reality gap = "surprise"

4 Reweight & update
Upweight surprising rollouts: GRPO, OPD, Combined

Co-evolution Shared parameters: predictor and policy co-evolve, and focus shifts to remaining blind spots

Calibration Cells & the Update Rule

Calibration cells (z × y)



Surprise-weighted update

Prospective-hindsight loss

$$\ell_{PH} = (1 + \alpha \cdot \xi) \cdot \ell_{base}$$

Surprise $\xi = \mathbb{1}[(z, y) \in \{OF, US\}]$
ξ is stop-gradient

α **Single hyperparameter α**
α = 0 recovers base method · default α = 1

M **Self-evaluator: same policy**
Sees only (s, a); ternary answer in \boxed{}; majority vote over M samples

+ **Only overhead**
M extra forward passes per step

~6% single-turn sync | ≈0% async multi-turn

Calibration Emerges as a Byproduct

$E[\ell_PH] = E[\ell_base] + \alpha \cdot R(\theta), \quad R(\theta) = E[\xi \cdot \ell_base] \geq 0$

Exact identity

$$R(\theta) = c(\theta) \cdot M(\theta)$$

M(θ) = Pr[OF or US]: miscalibration rate (= binary Brier score of the self-evaluator)

c(θ): mean base loss on surprising rollouts

R(θ) descent

- **↓M: calibrate belief** — align belief with outcome
- **↓c: fix hard cases** — fix miscalibrated rollouts

No explicit calibration loss → avoids the capability–calibration trade-off of calibration-augmented RL

1 Early

c and M both large

→ **Largest PH gains**

2 Mid

Both shrink as the self-model improves

3 Convergence

c, M → 0; R vanishes

PH shares the base method's fixed points

Caveat: a motivation for reduced miscalibration, not a formal guarantee.

Experimental Setup

🎯 Single-turn (verifiable)

- **Tasks:** Science Q&A (SciKnowEval), Tool Use (ToolAlpaca)
- **Rule-based verifier;** base loss SDPO; **M = 1**
- **Controls** (same 1+α multiplier, different selection): **Random reweight, Failure-only reweight**

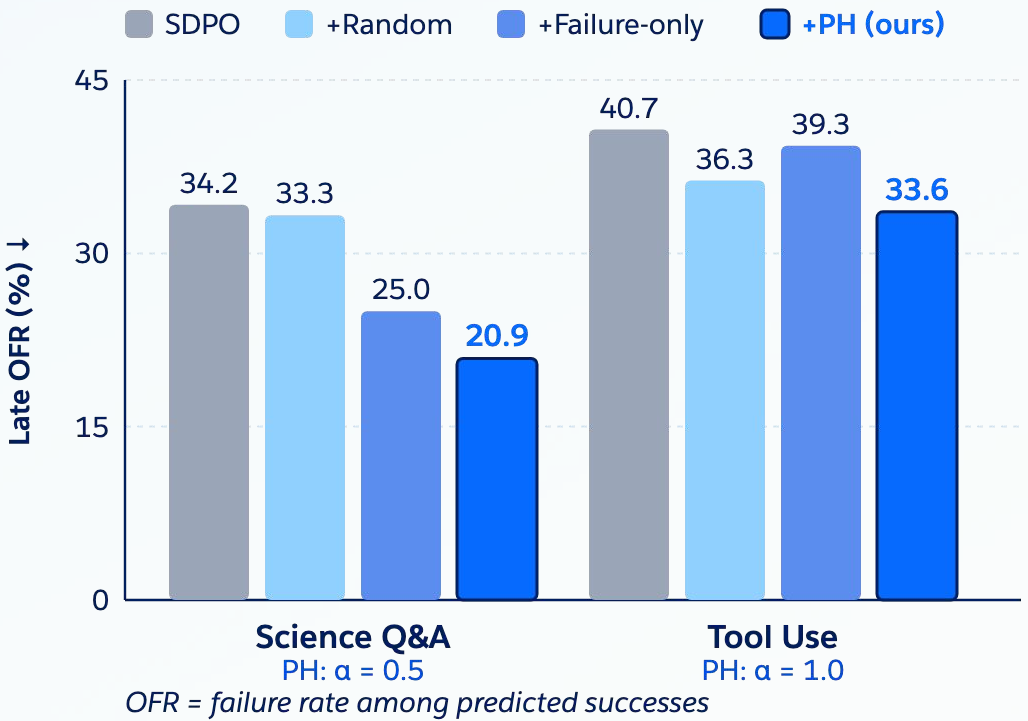
💬 Multi-turn (personal agent)

- **OpenClaw-RL** personal agent (math tutoring), **GPT-4.1** user simulator
- **LLM-based PRM judge**
- **PH** on GRPO, OPD (SDPO), Combined RL+OPD; **M = 3**

Models	**Policy = self-evaluator:** OLMo-3-7B-Instruct (shared weights, distinct prompts)		**Scale check:** GPT-OSS-20B	**3 seeds**
Metrics	**Capability** (mean@16 / judge score)	**OFR** = Pr[y=0 | z=+1]	**Surprise** = Pr[ξ=1]	**Prospective Acc.** = 1 − Surprise

Selection, Not Mass, Drives the Gain

Single-turn Late-training OFR (%, lower is better) across reweighting variants



Science Q&A
−13.3
pts OFR
34.2 → 20.9

Science Q&A
+2.7
pts avg acc.
63.0 → 65.7

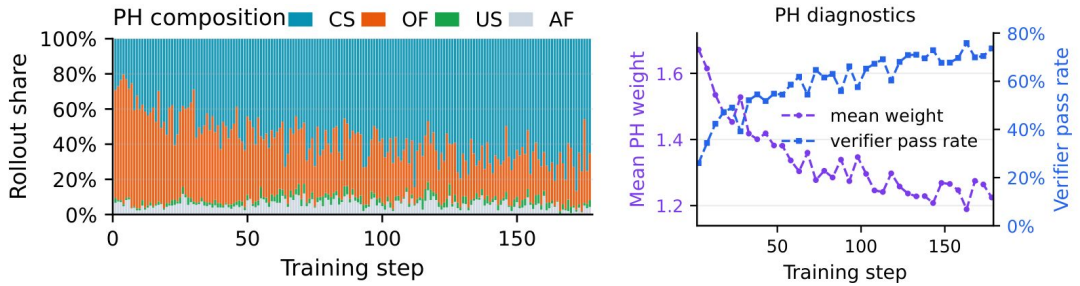
Takeaways

- **Random ≈ no gain** at equal weight: which rollouts matter, not how much.
- **Failure-only** recovers most OFR but trails on surprise (24.4 vs 21.8): misses underconfident successes.
- **Surprise tracks OFR:** 35.2 → 21.8 (Science); 41.6 → 33.5 (Tool Use).
- **Capability kept:** Science max 73.4 → 75.5; Tool Use avg/max 57.7/60.2 → 58.2/61.7, best jointly.

Self-Extinguishing Curriculum That Scales

Dynamics · Science Q&A, α=1

OF share ↓ · CS ↑ · mean PH weight ↓



Prospective accuracy

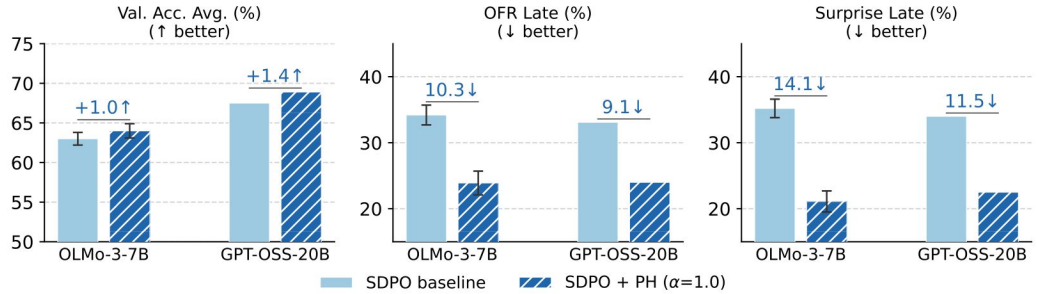
~30% → ~80%

vs. ~73% plateau for SDPO

No direct gradient to the predictor (stop-gradient ξ)

Scale · same α=1.0

SDPO → **SDPO+PH** · OLMo-3-7B · GPT-OSS-20B



Weight anneals as pass rate rises – no schedule

Gains hold at 7B and 20B

OLMo-3-7B

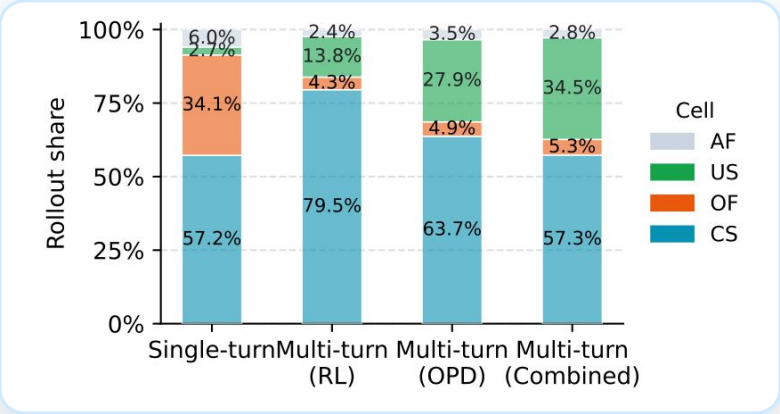
val. acc. **+1.0** · late OFR **−10.3**late surprise **−14.1**

GPT-OSS-20B

val. acc. **+1.4** · late OFR **−9.1**late surprise **−11.5**

Multi-Turn: Miscalibration Mode Flips

Dominant error flips



Single-turn: over-trust wrong actions (OF 34.1%, US 2.7%)

Multi-turn: under-trust correct ones (US 13.8 → 27.9 → 34.5%, RL < OPD < Combined; OF stays 4–5%)

PH (α=1.0) lifts all methods

Avg. judge score, base → +PH · small line = best score

Base	Avg score	Δ	OFR (%)
GRPO	68.6 → **75.8** best 77.7 → 82.9	+7.2	5.3 → **3.1**
OPD	54.0 → **69.3** best 65.6 → 75.2	+15.3	15.9 → **2.6**
Combined	61.2 → **65.1** best 73.5 → 78.5	+3.9	10.0 → **6.8**

Prospective acc. peaks **87.8%** (GRPO+PH) · **86.3%** (OPD+PH, α=0.5)

Caveat: α=2.0 gives lowest OFR (2.5%) but the predictor collapses toward constant z=−1

Same pattern holds with the GPT-5.4 user simulator

Takeaways & Outlook

1 Gap as signal

Prediction–reality disagreement is a **training resource**, not just a diagnostic.

2 One-line, plug-in

$(1 + \alpha \cdot \xi)$ reweighting on GRPO, OPD, and Combined; one hyperparameter α transfers across tasks, base methods, and scales.

3 Capability + calibration together

Calibration emerges as a byproduct: **no trade-off, no extra loss**; the weight self-extinguishes as the gap closes.

4 Limits & next steps

Limits: high α (≥ 2) can destabilize; relies on initial self-model quality. **Next:** longer horizons (SWE, terminal, GUI, research agents; current cap 48 turns).