Acquire, Repair, Preserve: A Diagnosis-Guided Post-Training Recipe for Small-Model Dialogue Game Agents

Nan Li · Utrecht University

Paper Model

Broad SFT brings most of the improvement; turn-local preference pairs target mechanically verifiable failures; OOD gains concentrate in related Wordle variants.

PUBLIC CLEMSCORE	CLOSED IN-DOMAIN	CLOSED OUT-OF-DOMAIN	STATIC (PUBLIC)
10.67 → 38.92	13.41 → 41.17†	3.72 → 7.88	44.24 → 44.14†

Official final evaluation (20 Aug 2026 snapshot) · Qwen3.5-2B base → ours

PART 1 · THE RECIPE — Acquire broadly, repair narrowly, preserve in weight space

DIAGNOSIS · REAL TRANSCRIPT

A valid explanation, then an invalid move

```
Model guess: slate
GM    guess_feedback: s<red> l<red> a<red> t<red> e<red>
Model guess: crane
GM    guess_feedback: c<red> r<green> a<red> n<green> e<red>
Model explanation: … I need to find a word with 'r' and 'n' in the correct
      positions, and avoid the letters 'c', 'a', and 'e'.
      guess: ringer        6 letters | uses e | r, n misplaced
GM    Error: The length of the guessed word is not 5. game_result = ABORT
```

Figure 1 Real transcript, broad-SFT checkpoint (public-dev instance 14, target *grind*). The model states the constraint, then breaks it. Such post-feedback failures are frequent, local to one turn, and mechanically checkable.

SETTING

We post-train **Qwen3.5-2B** with LoRA (*r* = 16, α = 32, all-linear) to play the clembench dialogue games of the LM Playschool Challenge.

SFT: LR 2e-4, batch 32, max length 2,048 · DPO (B, C): LR 5e-6, batch 16, max length 1,024, β = 0.1 / 0.2, frozen reference · all 1 epoch

$$\text{clemscore} = \text{mean Played} \times \text{mean Quality} / 100$$

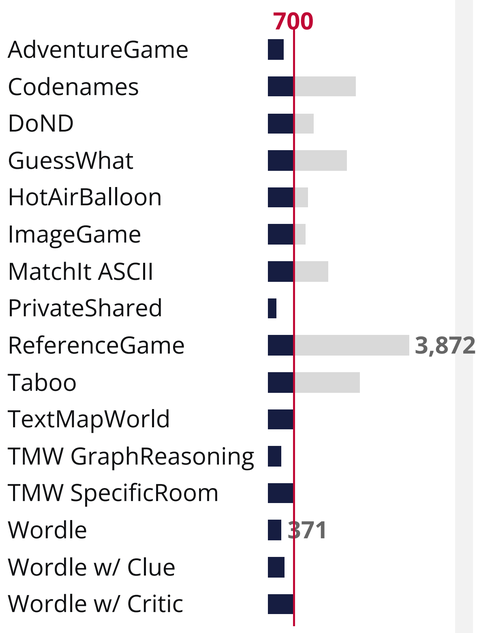
Played = % of episodes not aborted · **Quality** = task score on played episodes

TRAJECTORY · CLEMSCORE

Qwen3.5-2B 13.05 → **A Acquire** 43.85 → **B Repair** B1 46.83 → B2 45.52 → **C Refine** 48.52 → **D Preserve** 50.43* merge **48.70**

Figure 2 Training trajectory on our pinned local public-dev set (14 games) — not the official evaluation. B and C train on Wordle only. *Mean of two runs; the submitted merge scores 48.70.

A Acquire
broad SFT · all games



Data. Success-only, train split · 20,017 → 9,522 kept → 9,045 train (95/5)

B Repair
turn-local DPO · predefined errors

Source. Clean successful Wordle episodes (no invalid word, no repeat) · every reply from turn 2 that follows feedback.

h	… guess_feedback: c<red> r<green> a<red> n<green> e<red>		
y^+	guess: brink	*original reply*	
y^-	guess: crane	*repeat_last*	B1 1,905 pairs
	guess: brin	*bad_length*	
	guess: brinx	*invalid_word*	B2 1,906 pairs
	guess: brink guess: brink	*duplicate_guess_keyword*	

Schematic — not a training row.

WHY TURN-LOCAL

Pairs	Data	Clem (A = 43.85)	What broke
whole dialogue, unfiltered	3,973	**12.35**	prose before moves
whole dialogue, source-filtered	1,337	**25.76**	Wordle, Codenames abort

same *h*, next reply only → **Phase B**

Motivated the design; not a granularity-only ablation.

C Refine
branch DPO · model's own play

h shared history

→ next reply b_a → play full episode → R_a

→ next reply b_b → play full episode → R_b

pair next replies: best vs. worst branch

27 train-split Wordle instances · branch at rounds 1–3 · *R* = full-episode score

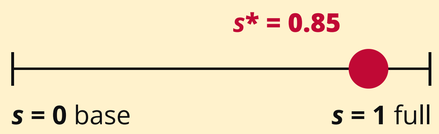
- **90** trajectories
- → **295** branch points
- → **77** pairs (73 train / 4 eval)

D Preserve
delta scaling · no gradient

Why. Every strong candidate paid a specialization tax on static scores.

$$W(s) = W_{\text{base}} + s \cdot \Delta W_{\text{LoRA}}$$
$$0 \le s \le 1$$
$$s^* = \operatorname{argmax}_s S(W(s))$$
$$\text{subject to}$$
$$C(W(s)) \ge C_{\min}$$

S = static, *C* = clemscore, both on public dev.



85% of the update, not of the capability.

PART 2 · RESULTS — What the recipe buys and where it stops

PLAYED × QUALITY · BASE → OURS

	PLAYED	QUALITY	CLEM
Public	27.22 → 78.63	39.20 → 49.50	**10.67 → 38.92**
Closed ID	47.32 → 82.62	28.33 → 49.83	**13.41 → 41.17†**
Closed OOD	35.19 → **33.08**	10.56 → 23.81	**3.72 → 7.88**

OOD Played falls below the base (red): the OOD gain is quality on fewer played episodes.

PUBLIC CLEMSCORE



Official baselines, no training.

PUBLIC STATIC COMPONENTS

PUBLIC STATIC · QUALITY	BASE	OURS	Δ
BBH	0.00†	30.47	**+30.47**
CLadder	48.51	46.53	**−1.98**
EQ-Bench	64.78	56.43	**−8.35**
IFEval	68.52	51.85	**−16.67**
MMLU-Pro	40.71	38.05	**−2.66**
Statscore	**44.24**	**44.14†**	

PER-GAME Δ EFFECTIVE SCORE · BASE → OURS

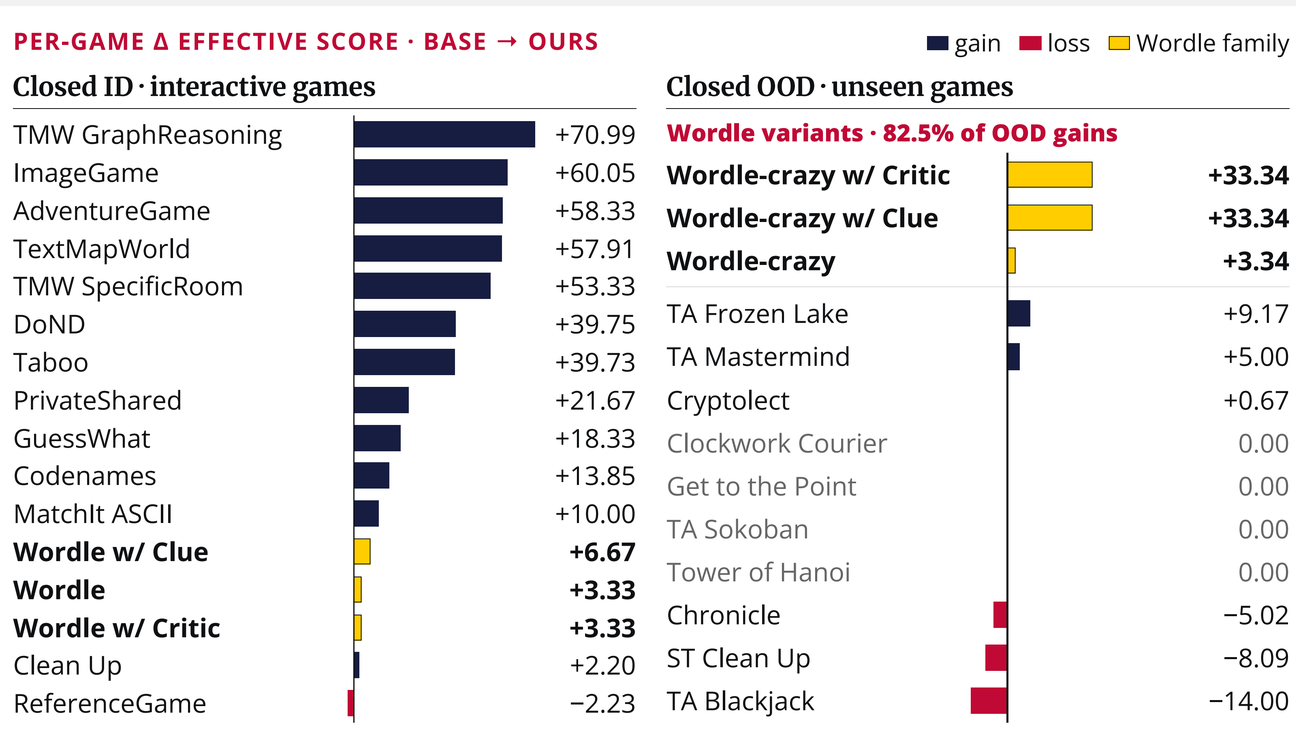


Figure 3 Effective = Played × Quality / 100; diagnostic, not averaged into clemscore. Closed-ID static configurations omitted.

The largest gains occur outside Wordle, the target of our repair stages; in-domain Wordle games end at 3–10 effective score, and OOD gains concentrate in Wordle variants.

† Paper numbers (snapshot `2dd5a533`, 20 Aug 2026). The organizers' re-scoring on 18 Sep 2026 (`8e0f470`) separated static tasks from closed ID and fixed BBH answer parsing (base BBH 0.00 → 10.94): closed ID 6.70 → 40.24; public Statscore 46.41 → 44.14; closed static 41.92 → 42.93. Public and OOD clemscores unchanged.

Nan Li · n.li@uu.nl · nanli.nl