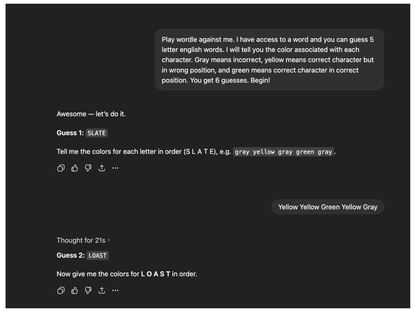
Policy Gradients and Information-Theoretic Wordle Solving

Andrew Sung [1] Ethan Hersch [1] Manav Kant [1]

[1]Department of Computer Science, Stanford University — CS 234: Reinforcement Learning

Problem

Wordle looks simple. Solving it with RL isn't.



- Action space: **2,315** guesses; horizon: **6** steps (sparse rewards)
- Observation: 5 letters $\times$ (position, letter, color) $\Rightarrow |S| \approx 26^{30} \cdot 3^{30}$

Background

Information Theory for Wordle

The optimal greedy strategy maximizes expected information gain per guess (Sanderson, 2022):

$$\mathbb{E}[\text{Info}] = \sum_x p(x) \cdot \log_2\left(\frac{1}{p(x)}\right) = H(X)$$

Each guess yields one of $3^5 = 243$ feedback patterns. Greedy entropy achieves **3.4–3.6** avg guesses — near-optimal but *cannot plan ahead*.

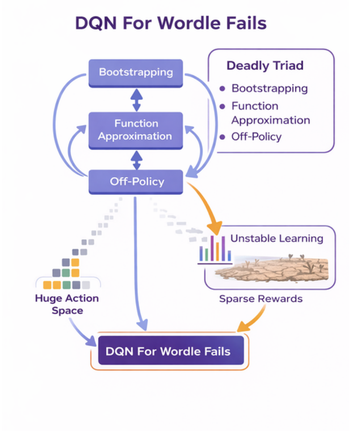
Key Insight

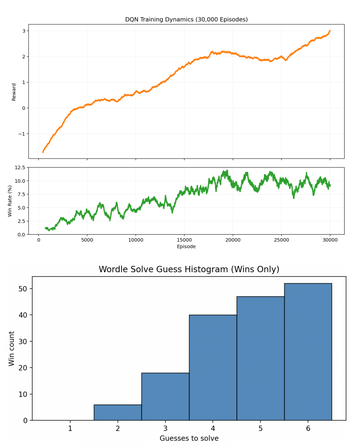
Use entropy as a *prior* to initialize RL, then learn *when to deviate* for multi-step planning gains.

Baselines

DQN: 3-layer MLP, ϵ-greedy. **Deadly Triad** $\rightarrow$ **11.5%** win rate.

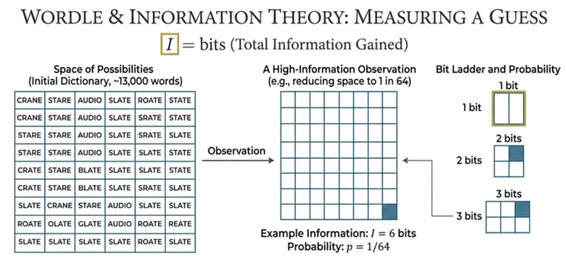
Vanilla PPO: letter-aware attention (5×26 features). Collapses to degenerate fixed point $\rightarrow$ **0%** win rate. (We are able to get to $\sim 99\%$ win-rate.)





Our Method: Entropy-Guided PPO

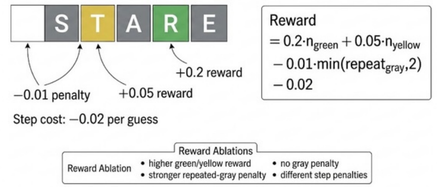
Our main contribution combines information-theoretic reasoning with learned RL corrections via PPO.



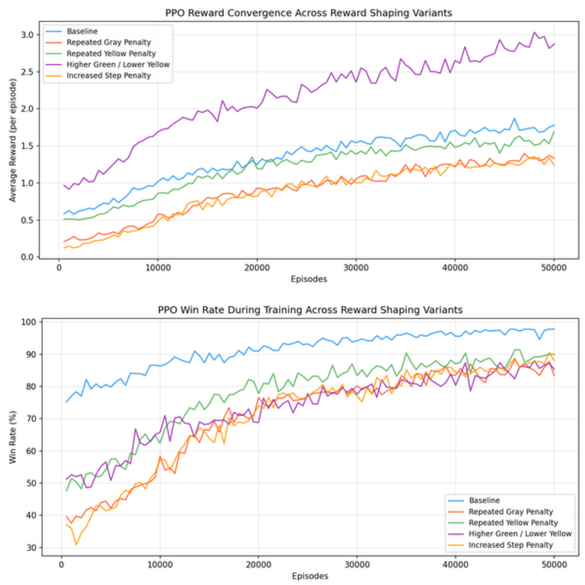
Intuition: start from an entropy-maximizing prior over guesses, then let PPO learn when to deviate for multi-step planning.

Reward Shaping

The default Wordle gym reward ($+10$ win, -1 per guess) is too sparse. We designed a granular per-letter reward:



Ablated 5 reward variants:

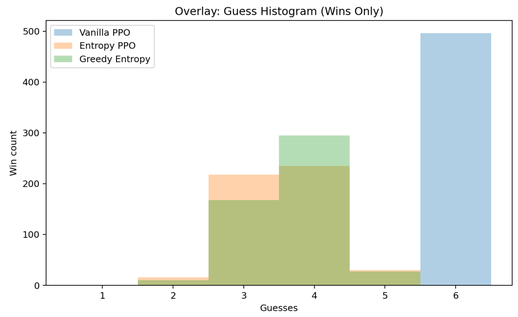


Experiments & Results

Evaluated over **500 games** on the full 2,315-word vocabulary.

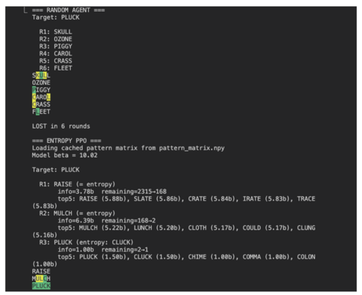
Table 1. Performance comparison across agents (500 games).

Agent	Win Rate	Avg Rounds
Random	0.2%	6.00
Vanilla PPO	99.6%	$\sim$ 6.00
Greedy Entropy	100.0%	3.61
Entropy PPO (Ours)	**100.0%**	**3.51**



- Entropy PPO: **100% win rate**, **3.51 avg rounds**
- Beats greedy entropy (3.61) and vanilla PPO ($\sim$6.00)

Analysis: When Does RL Deviate?



- Learned $\beta = 10.02$: heavy trust in entropy prior
- Agent deviates on **16%** of guesses, shifting games from 4–5 rounds into 3
- On PLUCK: greedy entropy guesses CLUCK; Entropy PPO guesses the answer directly

Conclusions

- DQN and vanilla PPO **fail** without domain guidance
- Entropy-guided PPO: **100% win rate, 3.51 avg rounds**
- Reward shaping critical for sparse-reward word games

References

[1] Sanderson, G. Solving Wordle using information theory. 3Blue1Brown, 2022.

[2] Kho, A. Wordle solver: Actor-critic deep RL, 2022.