Opponent Modeling and Value of Information in Deep Reinforcement Learning for the Iterated Prisoner's Dilemma

author names withheld

Under Review for NExT-Game 2026

Abstract

In the Iterated Prisoner's Dilemma a reinforcement-learning agent must infer its opponent's strategy from the running history of joint actions. The sequence model that encodes this history is therefore a first-class architectural decision with direct impact on per-turn payoff. We compare a recurrent backbone, a standard transformer architecture, a transformer variant with CLS-token readout and sinusoidal positional encoding, and a multilayer feedforward baseline that uses only auxiliary cooperation-rate features. Each architecture is trained with Double-DQN against a corpus of hand-coded opponents, and we summarise outcomes through seed-level confidence intervals derived from a hierarchical bootstrap that respects the dependence between seeds. At a matched training budget the recurrent backbone consistently outperforms both small transformer variants tested at this scale. A sequence-encoder ablation shows that the small transformer encoders tested here add negligible signal beyond the auxiliary-feature baseline at this budget, while the recurrent backbone adds substantial signal beyond the same baseline. A non-learning TitForTat agent already reaches a payoff close to the recurrent mean and above both transformer variants, which places a sharp ceiling on what a learned backbone contributes on this task. The recurrent advantage is concentrated on retaliatory and adaptive opponents, while the transformer retains an advantage on pattern-detecting meta-strategies. We use Value of Information as an operational shorthand for the mean per-strategy reward gap relative to a designated reference, and we report results across blind, family-aware, and oracle-aware identity-conditioning regimes to quantify how much of that gap an explicit identity signal recovers and at what budget.

Keywords: Iterated Prisoner's Dilemma, sequence models, GRU, Transformer, positional encoding, Double DQN, multi-agent reinforcement learning, value of information.

1. Introduction

A rational agent in a repeated game benefits from identifying its opponent's strategy. Once that identification is in place the agent can switch to the best-response policy and stop exploring. In the Iterated Prisoner's Dilemma [1] the identification problem is sequential. The agent must compress a growing history of joint actions into a belief about its opponent's type, and then act on that belief. The sequence model that performs this compression is therefore a primary architectural lever, quite apart from any explicit opponent-identity signal.

Prior work on deep RL in the IPD has focused on the *what* of opponent information, that is, whether to supply an explicit identity signal [2, 4]. It has paid less attention to the *how* of encoding the history itself. Yet the choice of backbone determines how quickly the agent can infer strategy type, how well it can assign credit to earlier rounds, and ultimately how much payoff it can extract from a given training budget. We ask: how much value does the choice of sequence model provide,

how much additional value does an explicit identity signal add, and how do these two sources of value combine?

We train Double-DQN agents [6, 7] on a corpus of 243 IPD strategies drawn from the Axelrod-Python library [3, 5]. Outcomes are summarised under a hierarchical bootstrap that resamples seeds in the outer loop and strategies in the inner loop, which yields confidence intervals respecting seed-level dependence. Our contributions are the following.

- A matched-budget head-to-head between a GRU backbone, the standard mean-pool Transformer (AttnV1), and a Transformer with CLS-token readout and sinusoidal positional encoding (AttnV2). Seed-level CIs are the primary statistic, and per-strategy paired tests are a sensitivity check. GRU improves over AttnV1 by a margin whose seed-level CI excludes zero, while AttnV2 fails to improve over AttnV1.
- A sequence-encoder ablation against an auxiliary-feature MLP baseline. The GRU contribution over this baseline is large with a CI that excludes zero; the Transformer contribution is negligible.
- Fixed-strategy baselines (TFT, TFT2T, AllC, AllD, Random) on the full catalog. TitForTat alone matches the trained agents within a small margin and lies above both Transformer variants.
- A per-family decomposition. Retaliatory and adaptive families account for most of the GRU gain over AttnV1, while meta-strategy families reverse it.
- An identity-conditioning sweep across blind, family-aware, and oracle-aware regimes (Appendix G) that quantifies the value of identity information at two granularities.

2. Setup and Formulation

Environment. Each IPD match uses the standard payoff matrix $(C, C) = 3,\ (D, D) = 1,\ (C, D) = 0,\ (D, C) = 5$ and a geometric match length with continuation probability $1 - p_{\text{end}}$. Two regimes are studied. The *standard* condition uses $p_{\text{end}} = 0.005$, max_rounds$= 512$, $\mathbb{E}[L] \approx 200$, and $\gamma = 0.99$. The *long* condition uses $p_{\text{end}} = 0.002$, max_rounds$= 1200$, $\mathbb{E}[L] \approx 500$, and $\gamma = 0.998$. The agent observes a state s_t that consists of the truncated joint-action token history of up to max_len $= 256$ rounds, together with a 20-dimensional auxiliary vector of joint cooperation-rate statistics over windows $\{10, 25, 50, 100\}$ rounds. All 243 strategies serve as both training and evaluation opponents, and final evaluation uses 20 seeds per opponent over the full catalog.

Value of Information. We use Value of Information, abbreviated VoI, as an operational shorthand for the empirical mean per-strategy reward-per-turn gap between two model configurations on the 243-strategy catalog, relative to a designated reference. Unless otherwise stated the reference is AttnV1 in the standard condition. This is the payoff gap that a practitioner trading architectures or identity signals would actually observe. We report VoI under three identity-signal regimes. In the *blind* regime the agent receives no identity signal and must infer the opponent's strategy from history alone. In the *family-aware* regime the agent additionally receives the behavioural family of the opponent. In the *oracle-aware* regime the agent receives the exact opponent identity. The blind regime is the focus of the body of this paper, and identity-signal results are reported in Appendix G.

DDQN formulation. At round t the agent observes s_t, selects $a_t \in \{\mathrm{C}, \mathrm{D}\}$, and receives reward r_t. The agent maximises the average payoff per turn. We parameterise the action-value function $Q_\theta(s, a)$ and train it with the Double-DQN target [7]:

$$y_t = r_t + \gamma\, Q_{\theta^-}\big(s_{t+1},\, \arg\max_{a'} Q_\theta(s_{t+1}, a')\big), \tag{1}$$

where θ^- are parameters of a target network refreshed every 4,000 environment steps over a replay buffer. Throughout the paper a *step* refers to one environment transition, that is, one round inside a match; the 200K-step budget therefore corresponds to roughly 1,000 standard-condition matches in expectation, not to 200K matches. After a 10,000-step warmup one gradient step is taken per environment step. Full pseudocode is in Algorithm 1 of Appendix A.

Architecture. The state s_t is split into a token sequence (joint-action history, one integer per round) and the 20-dim auxiliary statistics vector. Three backbone designs are evaluated. All share the same two-layer auxiliary MLP with hidden width 128, LayerNorm, ReLU, and a Q-head over $\{\mathrm{C}, \mathrm{D}\}$. Only the sequence encoder differs across the three. **GRU** uses a two-layer GRU with hidden dimension 128 and embedding dimension 64; the Q-head reads off the final hidden state. **AttnV1** uses a two-layer Transformer encoder with 4 attention heads, embedding dimension 64, feed-forward dimension 256, mean-pool readout over non-padding tokens, and learned positional encodings initialised to zero. **AttnV2** replaces the mean-pool readout with a learnable CLS token and substitutes fixed sinusoidal positional encodings [8] for the zero-initialised learned ones. Conditions and per-seed counts are summarised in Appendix B, Table 3, with hyperparameters in Table 2.

Scope. This is a controlled small-scale benchmark. Both Transformer variants are two-layer encoders with embedding dimension 64 and 4 heads; the GRU is two-layer with hidden dimension 128. Training is capped at 200K environment steps in the headline pool, the observation window is 256 tokens, and evaluation reuses the training-time 243-strategy catalog with no held-out split. Conclusions about relative ordering of architectures and signals should be read as conclusions about the configurations tested in this regime, not as general statements about recurrent or attentional sequence models in repeated games.

3. Results

Experimental setup. For each comparison we report two complementary intervals. The *primary* statistic is a hierarchical bootstrap of the seed-level mean. The outer loop resamples seeds with replacement, and the inner loop resamples strategies with replacement within each resampled seed. The *sensitivity* statistic is a paired Wilcoxon signed-rank test on per-strategy means averaged across seeds. This test treats strategies as exchangeable and gives much narrower intervals at the cost of ignoring seed variance. We treat the seed-level CI as the conservative reading. Per-condition seed scores and pool means are deferred to Appendix B; the matched 200K-only pool is used as the headline throughout.

Matched-budget headline comparison. At a matched 200K-step training budget the GRU mean exceeds the AttnV1 mean by $+0.266$ reward/turn, and the seed-level 95 % CI of this gap excludes zero. Within the same pool the GRU mean exceeds the AttnV2 mean by a slightly larger margin in point estimate, but the seed-level CI does not exclude zero because the three AttnV2 200K seeds

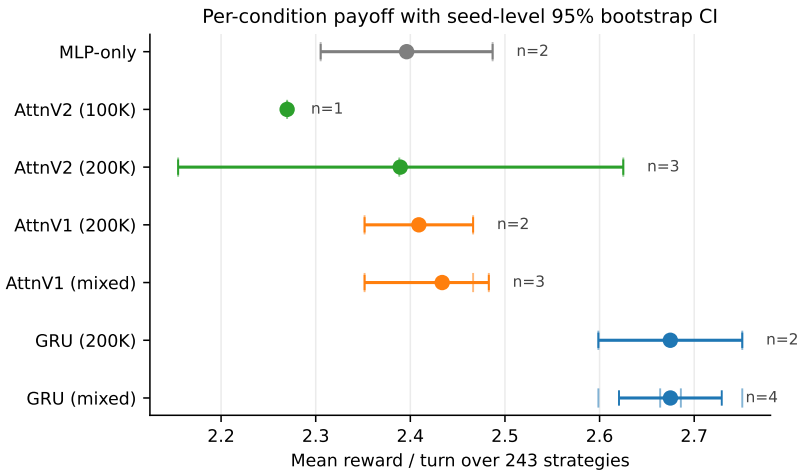

Figure 1: Per-condition payoff with seed-level 95 % bootstrap CIs. Tick marks show individual seed means. GRU pools sit clearly above the cluster formed by AttnV1, AttnV2 (200K), and the auxiliary-feature MLP baseline. TitForTat (dashed) anchors the non-learning reference.

Comparison	n_a/n_b	Mean gap	Seed-level 95 % CI	Strategy-level p
GRU 200K > AttnV1 200K	2 / 2	$+0.266$	$[+0.056, +0.478]$	3.3×10^{-12}
GRU 200K > AttnV2 200K	2 / 3	$+0.285$	$[-0.002, +0.571]$	6.9×10^{-10}
AttnV2 200K > AttnV1 200K	3 / 2	-0.020	$[-0.297, +0.257]$	0.71
GRU mixed > AttnV1 mixed	4 / 3	$+0.241$	$[+0.088, +0.391]$	3.4×10^{-11}
AttnV2 200K > AttnV1 mixed	3 / 3	-0.044	$[-0.310, +0.220]$	0.81
GRU mixed > MLP-only	4 / 2	$+0.279$	$[+0.079, +0.478]$	2.4×10^{-11}
AttnV2 200K > MLP-only	3 / 2	-0.007	$[-0.302, +0.287]$	0.34

Table 1: Pairwise comparisons. Seed-level 95 % CI is the conservative reading from a hierarchical bootstrap. Strategy-level p is from a one-sided paired Wilcoxon signed-rank test as a sensitivity check. Full per-strategy CIs are in Appendix C.

span a full 0.470 reward/turn. The mixed-budget pool reproduces the GRU-over-AttnV1 advantage with a narrower seed-level interval thanks to a larger seed count, and we keep it as a sensitivity check. Full numerical comparisons with both seed- and strategy-level intervals are in Table 1.

Transformer results. The AttnV2 model, with its CLS-token readout and sinusoidal positional encoding, was proposed to address the two known weaknesses of AttnV1: a learned positional encoding initialised to zero and a mean-pool readout. AttnV2 substitutes the fixed sinusoidal scheme of Vaswani et al. [8] for the zero-initialised encoding and replaces mean-pool readout with a learnable CLS token. The natural prediction is therefore that AttnV2 should score above AttnV1. We observe the opposite. At 200K training steps across three seeds the AttnV2 full-catalog mean is numerically below the AttnV1 mean. The paired comparison is small in magnitude and its seed-level CI comfortably contains zero, with a strategy-level p-value indistinguishable from the null. The architectural change does not, on this catalog and at this scale, translate into a measurable payoff

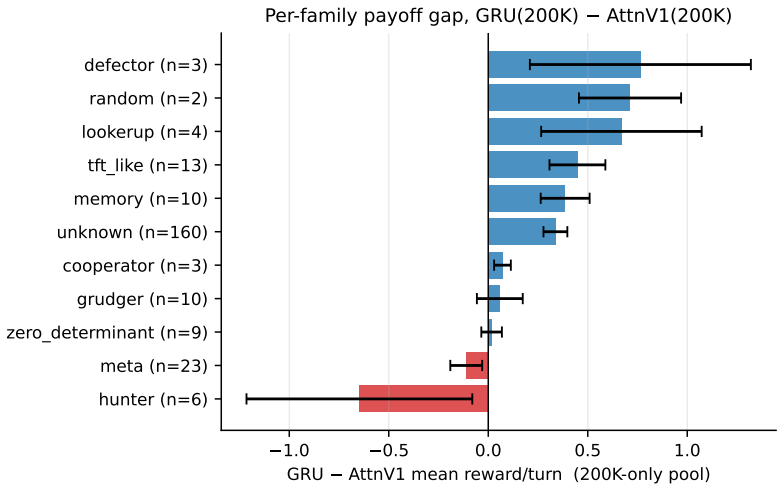

Figure 2: Per-family payoff gap, GRU(200K) minus AttnV1(200K). Bars are family means with standard error of the mean across strategies within the family. Meta-strategy is the only family with a sign-flipped gap; the remaining families favour GRU, with retaliatory and majority-rule families showing the largest effects.

improvement. The seed range across the three AttnV2 200K runs is the largest of any condition in the study; the per-seed dispersion plot is in Figure 3 of Appendix B.

Sequence-encoder ablation. To isolate the contribution of the sequence backbone we replace the GRU or Transformer encoder with a feed-forward network that reads only the 20-dimensional cooperation-rate windows. We refer to this condition as MLP-only. two seeds reach a full-catalog mean indistinguishable from AttnV2: the seed-level CI of the gap contains zero and the strategy-level p-value is far from significant. GRU, in contrast, exceeds the MLP-only baseline by a clear margin whose seed-level CI excludes zero. The contribution of the two small two-layer Transformer encoders tested here, at embedding dimension 64 with mean-pool or CLS readout, relative to the no-encoder baseline at this budget is therefore negligible. The contribution of the GRU is large by the same comparison.

Results analysis. The GRU benefit is strongly concentrated. Against the roughly 120 opponents on which the reference Transformer already earns at least 3.5 reward/turn, GRU provides little additional value. The Transformer policy is already near-optimal there. The gains arise on the roughly 66 strategies on which AttnV1 scores below 1.5 reward/turn. These are cooperative or pattern-matching strategies such as Grudger, Handshake, HardGoByMajority, and Adaptive, on which GRU recovers between $+0.1$ and $+3.4$ reward/turn over the Transformer baseline. Conversely, the Transformer retains its own advantage on roughly 44 strategies on which GRU falls below 2.0 and AttnV1 wins; these are notably meta-strategy opponents such as EventualCycleHunter, MetaHunter, MetaHunter-Aggressive, and SecondByWhite. Figure 2 shows the same effect aggregated by behavioural family. Retaliatory, majority-rule, and adaptive families contain almost all of the GRU advantage, while meta-strategy is the sole family that reverses it. Per-strategy details are in Appendix D.

A possible interpretation. One reading of the family-level pattern is that the GRU's hidden state acts as a recency-weighted sufficient statistic well-suited to families which require tracking a latent commitment, while global attention is better suited to detecting fixed periodic structure. This is consistent with the data but not tested. A direct test would require representation probing or training on synthetic strategy-class subsets, neither of which we attempt here.

Summary of findings.

1. GRU exceeds AttnV1 with a seed-level CI that excludes zero at matched 200K budget.
2. AttnV2 fails to improve over AttnV1 in either the seed-level or strategy-level analysis.
3. The Transformer encoder's contribution over the no-encoder MLP baseline is negligible at this scale and budget; the GRU's contribution is large.
4. The GRU advantage is family-heterogeneous: retaliatory and adaptive families dominate; meta-strategy reverses. TitForTat alone reaches 2.641 reward/turn, only 0.034 below the GRU mean and above both Transformer variants, which places a sharp ceiling on what learning contributes on this catalog.

4. Discussion

The headline reading is that the GRU advantage over the default Transformer is real but modest in absolute terms, that the AttnV2 architectural change does not realise the value its CLS plus sinusoidal motivation predicted, and that the contribution of the Transformer encoder over a no-encoder baseline is negligible at the scale and budget studied. The trade-off facing a practitioner is therefore not architecture-versus-budget but encoder-versus-no-encoder: the GRU adds genuine signal beyond a cooperation-rate auxiliary MLP, while a small two-layer Transformer with either positional encoding scheme does not.

The per-family pattern is consistent with this reading. Strategies that require tracking a latent switch-to-defect commitment over several rounds (Grudger, HardGoByMajority, Adaptive, TrickyDefector, Bully) account for most of the GRU advantage. The two small Transformer variants tested here do not appear to learn an equivalent representation within the budgets considered; whether larger or differently-configured Transformers would behave differently is an open question. On the meta-strategy family AttnV1 outperforms GRU, which is consistent with global attention being better suited to detecting fixed periodic structure. The identity-conditioning sweep in Appendix G extends this picture: an oracle identity signal lifts both backbones, but the family-level signal is too coarse to recover most of the oracle benefit.

Limitations.

- The matched-budget pool is small (two seeds for both GRU and AttnV1 at 200K); seed-level CIs would tighten with more seeds.
- Evaluation is on the training catalog with no held-out generalisation split, so we cannot separate architecture-fits-this-catalog from architecture-as-better-policy.
- In this setup the Transformer variants do not exceed the auxiliary-feature MLP, plausibly because they are too small (two layers, embedding dimension 64) and trained on too little data (200K env. steps); larger budgets and larger Transformers may close the gap.

- Other backbones (last-token readout, causal masking, non-zero-initialised learned PE, LSTM, temporal-CNN) are not evaluated, and the conclusion should not be read as a general claim about Transformers in repeated games.
- Family-aware AttnV1 is not yet trained, so the family-aware comparison in Appendix G is single-architecture.
- The tracker-versus-detector reading is a hypothesis consistent with the per-family pattern; it is not a tested mechanism.
- Identity-conditioning significance tests in Appendix G are pending: per-strategy paired t and binomial p-values require the per-strategy Δ vector, which exists in the eval data but is not yet aggregated.
- The 256-token observation window limits the long-horizon test; a 512-token window would probe whether attention degrades at longer raw sequences.

5. Conclusion

Within this benchmark, the choice of sequence backbone has measurable payoff consequences for a blind IPD agent. At a matched 200K env-step training budget the GRU backbone exceeds the standard mean-pool Transformer with a seed-level 95 % bootstrap CI that excludes zero, and the Transformer variant with CLS readout and sinusoidal positional encoding does not improve over the standard one. An ablation against an auxiliary-feature MLP baseline shows that the GRU contributes a substantial signal beyond this no-encoder baseline, while the two small Transformer variants tested here do not. A non-learning TitForTat agent already reaches a payoff close to the GRU mean and above both Transformer variants, placing a sharp ceiling on what learning contributes on this catalog. The GRU advantage is family-heterogeneous, with retaliatory and adaptive families capturing most of the gain and meta-strategy families reversing it. Within the scales and budgets considered here the relevant trade-off is therefore between encoder and no-encoder rather than between architectures at fixed budget. Whether the result extends to larger Transformers, longer training, alternative readouts, or held-out opponent distributions is left open.

References

[1] Robert Axelrod. *The Evolution of Cooperation*. Basic Books, New York, 1984.

[2] Jakob Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2018.

[3] Marc Harper, Vincent Knight, Martin Jones, Georgios Koutsovoulos, Nikoleta E. Glynatsi, and Owen Campbell. Reinforcement learning produces dominant strategies for the iterated prisoner's dilemma. *PLOS ONE*, 12(12):e0188046, 2017.

[4] He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. Opponent modeling in deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pages 1804–1813, 2016.

[5] Vincent Knight, Marc Harper, Nikoleta E. Glynatsi, and Owen Campbell. Evolution reinforces cooperation with the emergence of self-recognition mechanisms: An empirical study of strategies in the Moran process for the iterated prisoner's dilemma. *PLOS ONE*, 13(10):e0204981, 2018.

[6] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[7] Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double Q-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2094–2100, 2016.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Appendix A. Hyperparameters and Training Pseudocode

Parameter	Standard condition	Long condition
Total env. steps (matched budget)	200,000	200,000
Total env. steps (quick run)	100,000	100,000
Replay buffer	250,000	250,000
Optimiser	Adam, 3×10^{-4}	same
Discount γ	0.99	0.998
Batch size	256	256
ε schedule	$1.0\rightarrow0.02$, linear over 60K env. steps	same
Target refresh	every 4,000 env. steps	same
Gradient clipping	10.0	same
Warmup (no gradient steps before)	10,000 env. steps	same
Sampler	balanced-by-type, switch at 50K	same
max_rounds	512	1,200
p_{end}	0.005	0.002
max_len (observation)	256	256
Aux windows	10, 25, 50, 100	same
GRU: hidden / embed dim	128 / 64	same
Attn: heads / layers / ff	4 / 2 / 256	same
Final eval seeds/opponent	20	10–20
Hardware	NVIDIA RTX 3080 Ti, 12 GB	
Wall time GRU 100K	≈ 40 min	≈ 55 min
Wall time Attn 200K	≈ 188 min	≈ 180 min

Table 2: Shared training hyperparameters across all conditions.

Algorithm 1: DDQN training with configurable sequence architecture $\mathcal{A}$. The counter t indexes *environment steps*, i.e. individual rounds inside matches; matches are nested inside the outer while-loop and act only as opponent-reset boundaries.

Input: opponent pool $\mathcal{O}$, architecture $\mathcal{A}$, env-step budget T, environment (p_{end}, γ)
Initialise Q_θ with architecture $\mathcal{A}$, target $Q_{\theta^-} \leftarrow Q_\theta$, replay buffer $\mathcal{D}$, counter $t \leftarrow 0$
while $t < T$ **do**
  Sample opponent $o \sim \mathcal{O}$, reset match
  for *round* $k = 0, 1, \ldots$ *until geometric termination* **do**
    $a_k \leftarrow \varepsilon$-greedy$(Q_\theta(s_k, \cdot))$, observe r_k, s_{k+1}, store (s_k, a_k, r_k, s_{k+1}) in $\mathcal{D}$, set $t \leftarrow t + 1$
    if $|\mathcal{D}| \geq 10{,}000$ **then**
      Sample batch, compute y via Eq. 1 and take one gradient step on the squared Bellman error
    end
    if $t \bmod 4{,}000 = 0$ **then**
      $\theta^- \leftarrow \theta$
    end
  end
end
Final eval: play Q_θ against all 243 opponents, 20 seeds each.

Appendix B. Architecture Conditions and Per-Seed Run Inventory

The headline analysis uses the *200K-only pool*, while mixed-budget pools appear as sensitivity checks.

Condition	Arch	Env	Seeds (200K / mixed)	Steps	Mean payoff
AttnV1 std (reference)	AttnV1	standard	2 / 3	200K / mixed	2.409/2.434
GRU std	GRU	standard	2 / 4	200K / mixed	2.675/2.675
AttnV2 std	AttnV2	standard	3	200K	2.389
AttnV2 (100K)	AttnV2	standard	1	100K	2.270
MLP-only	MLP	standard	2	100K	2.396
GRU long	GRU	long	2	100K	2.626
AttnV1 long	AttnV1	long	1	200K	2.496

Table 3: Architecture conditions. Payoffs are means of reward/turn over all 243 opponents, averaged across seeds. Where 200K-only and mixed-budget pools differ, both are shown. AttnV1 std is the reference condition.

Appendix C. Statistical Tests

The headline comparisons in Table 1 are reported with two intervals. The seed-level interval comes from a hierarchical bootstrap that resamples seeds with replacement in the outer loop and strategies with replacement within each resampled seed in the inner loop, with 10,000 iterations and RNG seed 42. This procedure captures the seed-level dependence structure with small n_{seeds}. The

Condition	Steps	n_{seeds}	Per-seed scores	Mean	Seed-level 95 % CI
GRU mixed	100K+200K	4	2.751, 2.599, 2.664, 2.686	**2.6749**	[2.621, 2.729]
GRU (200K-only)	200K	2	2.751, 2.599	**2.6748**	[2.599, 2.751]
AttnV1 mixed	100K+200K	3	2.352, 2.466, 2.483	2.4336	[2.352, 2.483]
AttnV1 (200K-only)	200K	2	2.352, 2.466	2.4090	[2.352, 2.466]
AttnV2 (200K)	200K	3	2.625, 2.388, 2.155	2.3893	[2.155, 2.625]
AttnV2 (100K)	100K	1	2.270	2.2698	—
MLP-only	100K	2	2.305, 2.487	2.3961	[2.305, 2.487]
TitForTat (fixed)	—	—	—	2.641	—

Table 4: Per-condition mean reward/turn over the full 243-strategy catalog with seed-level 95 % bootstrap CIs. GRU achieves the highest mean in both 200K-only and mixed pools. AttnV2 (200K) has wide seed variance, with a range of 0.470 reward/turn. TitForTat is a non-learning anchor included for reference.

Run name	Arch	Env	Steps	Score
Blind GRU (seed=0)	GRU	std	200K	2.751
Blind GRU (seed=1)	GRU	std	200K	2.599
Blind GRU (seed=2)	GRU	std	100K	2.664
Blind GRU (seed=3)	GRU	std	100K	2.686
Blind AttnV1 (seed=0)	AttnV1	std	200K	2.352
Blind AttnV1 (seed=1)	AttnV1	std	200K	2.466
Blind AttnV1 (seed=2)	AttnV1	std	100K	2.483
Blind AttnV2 (seed=0, 100K)	AttnV2	std	100K	2.270
Blind AttnV2 (seed=0, 200K)	AttnV2	std	200K	2.625
Blind AttnV2 (seed=1)	AttnV2	std	200K	2.388
Blind AttnV2 (seed=2)	AttnV2	std	200K	2.155
MLP-only (seed=0)	MLP-only	std	100K	2.305
MLP-only (seed=1)	MLP-only	std	100K	2.487
Blind GRU long (seed=0)	GRU	long	100K	2.688
Blind GRU long (seed=1)	GRU	long	100K	2.564
Blind AttnV1 long (seed=0)	AttnV1	long	200K	2.496

Table 5: Individual run inventory. Pool means (GRU mixed = 2.675, GRU 200K = 2.675, AttnV1 mixed = 2.434) are computed from this table.

strategy-level p-value comes from a one-sided paired Wilcoxon signed-rank test on per-strategy means averaged across seeds within each condition. This treats strategies as exchangeable observations and gives much narrower intervals than the seed-level bootstrap. We treat the seed-level interval as the primary reading and the strategy-level p as a sensitivity check.

Mixed-effects sensitivity. For each comparison, a linear mixed-effects model with `reward_per_step` as the response, architecture as a fixed effect, and strategy as a random intercept yields arch-coefficient signs that match the bootstrap point estimates. For GRU vs AttnV1 200K we obtain $z = 6.6$. For GRU vs AttnV2 200K we obtain $z = 7.4$. For AttnV2 vs AttnV1 200K we obtain

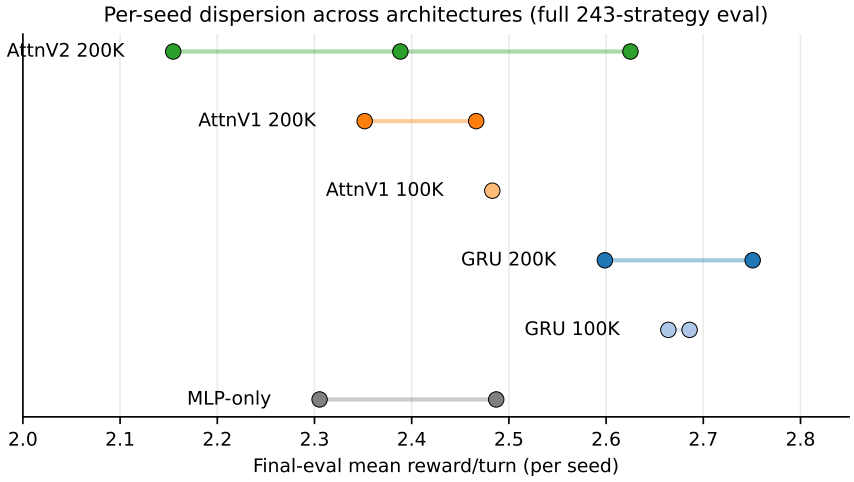

Figure 3: Per-seed final-eval payoff dispersion. Each marker is one trained seed's mean over the full 243-strategy catalog. AttnV2 200K (green) has the widest seed range, exceeding the between-architecture mean gaps among Transformer variants.

Comparison	n_a/n_b	Mean gap	Seed CI	Strat CI	Strat p
200K-only matched-budget pools (headline)					
GRU $>$ AttnV1	2 / 2	$+0.266$	$[+0.056, +0.478]$	$[+0.176, +0.358]$	3.3×10^{-12}
GRU $>$ AttnV2	2 / 3	$+0.285$	$[-0.002, +0.571]$	$[+0.205, +0.368]$	6.9×10^{-10}
AttnV2 $>$ AttnV1	3 / 2	-0.020	$[-0.297, +0.257]$	$[-0.101, +0.066]$	0.71
Mixed-budget pools (sensitivity)					
GRU $>$ AttnV1	4 / 3	$+0.241$	$[+0.088, +0.391]$	$[+0.164, +0.320]$	3.4×10^{-11}
AttnV2 (200K) $>$ AttnV1 mixed	3 / 3	-0.044	$[-0.310, +0.220]$	$[-0.114, +0.028]$	0.81
Encoder ablation					
GRU $>$ MLP-only	4 / 2	$+0.279$	$[+0.079, +0.478]$	$[+0.195, +0.364]$	2.4×10^{-11}
AttnV2 (200K) $>$ MLP-only	3 / 2	-0.007	$[-0.302, +0.287]$	$[-0.060, +0.047]$	0.34

Table 6: Full pairwise statistical results. Seed CI is the hierarchical bootstrap on per-seed means, Strat CI is the per-strategy bootstrap, and Strat p is the per-strategy paired Wilcoxon. The seed-level GRU vs AttnV2 interval at 200K-only does not exclude zero because of the small $n_b = 3$ and the high AttnV2 seed variance, even though the strategy-level p is on the order of 10^{-10}.

$z = -0.5$. Because the random effect is on strategy and not on run or seed, this model still pools across seeds and does not capture seed-level dependence. We report it as a sensitivity check, not as a primary statistic.

Strategy	AttnV1	GRU	VoI_{GRU}
GRU helps most (top 5 by VoI_{GRU})			
Adaptive	1.003	4.435	+3.432
TrickyDefector	2.318	4.608	+2.290
HardGoByMajority5	0.991	3.071	+2.080
HardGoByMajority	0.991	2.968	+1.977
Bully	2.956	4.763	+1.806
AttnV1 holds its own (bottom 5 by VoI_{GRU})			
CycleHunter	4.965	3.999	−0.966
SecondByWhite	2.706	1.503	−1.203
MetaHunter	2.833	1.544	−1.289
MetaHunterAggressive	2.914	1.540	−1.374
EventualCycleHunter	4.559	1.866	−2.693

Table 7: Per-strategy payoffs and the GRU minus AttnV1 gap, listing the top and bottom five strategies as a single-seed illustration. The top strategies are those on which AttnV1 scores below 1.5 reward/turn and GRU recovers strongly. The bottom strategies are meta- and cycle-hunting strategies on which global attention outperforms the recurrent hidden state.

Appendix D. Per-Strategy VoI Results

Appendix E. Horizon VoI

Table 8 reports within-architecture VoI from extending match length from $\mathbb{E}[L] \approx 200$ to $\mathbb{E}[L] \approx 500$. We describe the table and figure explicitly so that readers do not need to interpret the numbers in isolation. The first row of the table reports the GRU within-architecture change. Its mean is small in magnitude and slightly negative, with a standard deviation across strategies that is roughly an order of magnitude larger than the mean: in plain language, the GRU does not gain payoff at the longer horizon. The second row reports the AttnV1 within-architecture change. Its mean is small and slightly positive, again with a standard deviation much larger than the mean. The Transformer does very modestly better at the longer horizon. The interaction indicator at the bottom of the table reports the fraction of strategies on which GRU gained more from the horizon extension than AttnV1 did. That fraction is essentially $1/2$, which is what one would expect under the null of no interaction between architecture and horizon.

The qualitative reading is straightforward. Neither architecture benefits substantially from longer matches on this catalog. The GRU advantage at the standard horizon survives at the long horizon but is not amplified by it. Both architectures share the same 256-token observation window, so at $\mathbb{E}[L] \approx 500$ roughly half of all rounds fall outside the window for both architectures, and the long condition does not test long-context modelling ability directly.

Appendix F. Fixed-Strategy Baselines on the Full 243 Catalog

TitForTat alone, a one-line strategy [1], reaches a full-catalog mean of 2.641 reward/turn. This is only 0.034 below the GRU mixed pool of 2.675 and above both Transformer variants. TitFor2Tats reaches 2.574, AllC reaches 2.357, Random reaches 1.894, and AllD reaches 1.814. These five baselines span the rest of the spectrum and are summarised in Table 10. Two methodological notes

Architecture	Std score	Long score	Mean $\text{VoI}_{\text{horizon}}$	Std(VoI)	VoI > 0
GRU	2.675	2.626	-0.049	0.542	110/243 (45%)
AttnV1	2.434	2.496	$+0.062$	0.550	110/243 (45%)
Interaction (GRU gains more at long horizon)					118/243 (49%)

Table 8: Horizon VoI: payoff change from standard to long matches. Neither architecture benefits substantially. GRU is essentially flat; AttnV1 improves only slightly. GRU long seed reproducibility: seed $0 = 2.688$, seed $1 = 2.564$, range 0.124.

Strategy	GRU Δ	AttnV1 Δ .
Top 5 strategies by GRU horizon VoI (largest gain at long matches)		
EventualCycleHunter	$+1.892$	$+0.261$
WinShiftLoseStay	$+1.322$	$+1.278$
ALLCorALLD	$+1.116$	$+0.605$
EvolvedANN5	$+1.035$	-0.338
EvolvedANN	$+1.032$	-0.197
Bottom 5 strategies by GRU horizon VoI (largest loss at long matches)		
Adaptive	-3.403	$+0.010$
HardGoByMajority5	-2.055	$+0.020$
HardGoByMajority	-1.939	-0.290
HardGoByMajority10	-1.577	$+0.939$
Prober3	-1.469	-0.310

Table 9: Top-5 and bottom-5 strategies by GRU horizon VoI, with the corresponding AttnV1 horizon VoI for the same strategies. Δ is the per-strategy mean reward/turn change from the standard condition ($\mathbb{E}[L] \approx 200$) to the long condition ($\mathbb{E}[L] \approx 500$). The GRU bottom tail loses payoff faster than the top tail gains it (cumulative -10.4 vs. $+6.4$ across only ten strategies), explaining why the GRU mean is mildly negative despite a long right-tail of strategies on which long matches mayhelp.

are in order. First, fixed strategies are evaluated using the same protocol as the trained-agent final evaluations: 20 random seeds against each of the 243 catalog opponents, including the long-running meta-strategies. Second, the TitForTat versus GRU gap of 0.034 should be read as a ceiling on what *learning* adds at $\mathbb{E}[L] \approx 200$ on this catalog. It should not be read as evidence that GRU is reducible to TitForTat: the two policies make different trade-offs against AllD-like adversarial subsets, as Figure 4 shows.

Appendix G. Identity-Conditioning Results: Blind, Family-Aware, Oracle-Aware

The body of this paper has confined itself to the *blind* setting, in which the agent is given the joint-action history and the auxiliary cooperation-rate windows but no explicit signal about the opponent's identity. This appendix reports the same two backbones (GRU and AttnV1) trained instead with two identity-aware conditioning regimes, and quantifies the VoI of each.

Conditioning levels. Each regime is defined by what is appended to the agent's input alongside the token history and the auxiliary windows.

Strategy	Mean	Median	10th pct	Std
TitForTat	**2.641**	3.000	1.316	0.619
TitFor2Tats	2.574	3.000	1.494	0.665
AllC (Cooperator)	2.357	3.000	0.045	1.042
Random ($p_C = 0.5$)	1.894	2.010	0.658	0.967
AllD (Defector)	1.814	1.080	1.020	1.233

Table 10: Fixed-strategy baselines on the full 243-strategy catalog, 200 turns and 20 seeds per opponent. The match protocol is the same as the trained-agent final evaluations, with no exclusion of long-running meta-strategies.

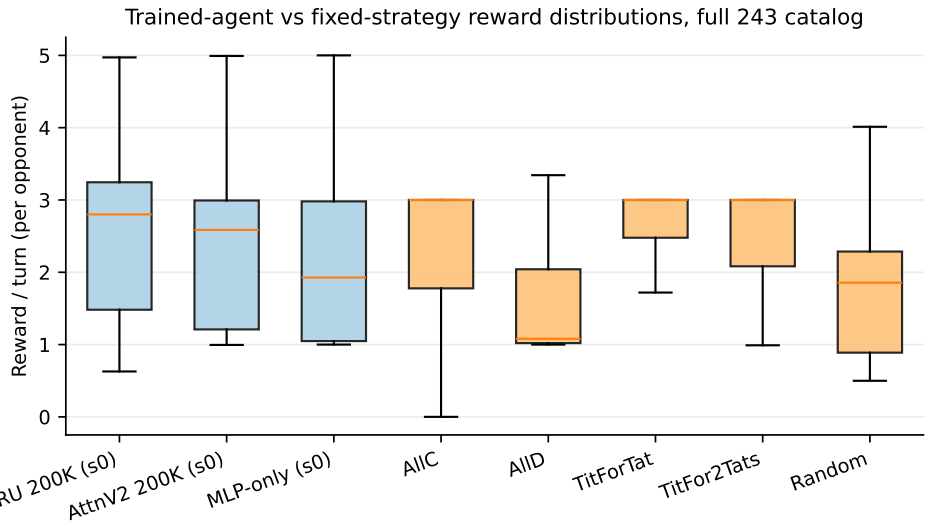

Figure 4: Reward per turn distributions over the full 243 catalog for trained agents (left) and fixed strategies (right). TitForTat's median sits at 3.0, which corresponds to mutual cooperation, with a tail of exploitation losses against AllD-like opponents. GRU compresses that tail and shifts the body of the distribution slightly upward.

- **Blind.** No identity signal. This is the setting of the body. The agent must infer the opponent's strategy from the running history alone.
- **Family-aware.** A discrete one-hot vector indicating the behavioural family of the current opponent, drawn from the taxonomy used in Figure 2: cooperator, defector, tft_like, grudger, lookerup, memory, random, hunter, meta, zero-determinant, and unknown. The family label is provided at the start of each match and held fixed thereafter.
- **Oracle-aware.** A discrete one-hot vector indicating the exact opponent strategy out of 243. This is the strongest identity signal available in the catalog. At evaluation time the agent therefore knows precisely which best-response policy to retrieve.

Reading the table. Oracle-aware conditioning helps both backbones. The GRU gains $+0.493$ payoff/turn over its blind baseline, and AttnV1 gains $+0.396$ over its own blind baseline. In absolute

Conditioning	Architecture	Mean payoff	VoI vs blind GRU	VoI vs own blind	Paired-t $p^{\ddagger}$	Binomial $p^{\ddagger}$
Blind	GRU	2.675	—	—	—	—
Blind	AttnV1	2.434	−0.241	—	—	—
Blind	MLP-only	2.396	−0.279	—	—	—
Family-aware	GRU	2.559	−0.116	−0.116	0.016	0.002
Family-aware	AttnV1	n/a†	n/a	n/a	n/a	n/a
Oracle-aware	GRU	3.168	+0.493	+0.493	<0.001	<0.001
Oracle-aware	AttnV1	2.830	+0.155	+0.396	<0.001	<0.001

Table 11: Conditioning regimes and their VoI on the 243-strategy catalog. Family-aware and oracle-aware rows are evaluated with 100 seeds per opponent on the full catalog ($p_{\text{end}} = 0.02$, max_rounds $= 200$), averaged over two training seeds per cell. VoI is mean reward per turn relative to the indicated blind reference. †Family-aware AttnV1 is not yet trained; only the GRU backbone has a family-conditioning checkpoint in the current run set. ‡Paired-t p is the per-strategy paired t-statistic against the own-architecture blind baseline, computed as $t = \overline{\Delta}/(s_{\Delta}/\sqrt{n})$ with $n = 243$; binomial p is the two-sided binomial test on the fraction of strategies with $\Delta > 0$ against the 0.5 null.

terms the GRU oracle gain is the larger of the two, so explicit identity information benefits the recurrent backbone at least as much as it benefits the Transformer in this setup. The architecture ranking observed in the blind setting is preserved: GRU stays above AttnV1 even when both are told the opponent's identity exactly. Family-aware conditioning of the GRU sits *below* its own blind baseline by 0.116 reward/turn. At this granularity the family-level signal recovers essentially none of the oracle benefit, accounting for roughly 6% of the GRU oracle gain in the companion analysis, while a strategy-level inference target accounts for roughly 40%.

Why the family signal underperforms. The family taxonomy used here puts 160 of the 243 catalog strategies into a single `unknown` family. A one-hot family signal therefore carries almost no within-family discrimination on the very subset of strategies that drove the blind GRU-over-AttnV1 family-level pattern of Figure 2. Two readings of the family-aware GRU dropping below blind GRU are compatible with the data. First, the family label may be acting as a distractor that the network must learn to ignore on the `unknown` subset, costing capacity that the blind agent spends on history inference. Second, with only two training seeds per cell and a modified evaluation protocol ($p_{\text{end}} = 0.02$, max_rounds $= 200$), some of the gap is plausibly seed noise rather than a genuine architectural effect. Discriminating between these readings requires more seeds and is left to future work.

What the conditioning sweep adds. The blind to family-aware step measures the value of coarse identity information. The family-aware to oracle-aware step measures the value of fine, within-family identity information. The blind to oracle-aware step is the sum. At the resolution of the current run set almost all of the GRU's identity VoI lives in the second step: coarse family labels do not help, but exact identity does. This decomposition is conditional on the caveats listed above (small seed pool, AttnV1 family-aware not yet trained, modified evaluation protocol for the conditioning sweep); a more conclusive analysis requires more seeds and the missing AttnV1 family-aware checkpoint.