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

Nazarii Tkach n.tkach@ukma.edu.ua
National University of Kyiv-Mohyla Academy, Kyiv, Ukraine

Oleksii Ignatenko o.ignatenko@ucu.edu.ua
Ukrainian Catholic University, Lviv, Ukraine

## 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 report architecture and budget comparisons as a mean per-strategy reward delta ($\bar{\Delta}$), and reserve Value of Information for the identity channel: the payoff gain from revealing the opponent's identity to a fixed architecture, which under a learned (hence only approximately optimal) policy is a lower bound on the expected value of perfect information. Across blind, family-aware, and oracle-aware regimes we quantify how much of the achievable gain an explicit identity signal recovers, and at what granularity and 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.

© N. Tkach & O. Ignatenko.

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 H) 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.

Two payoff differences: $\bar{\Delta}$ and VoI. We separate two distinct quantities that both have units of reward per turn. The first is the mean per-strategy reward delta $\bar{\Delta}$: the empirical mean over the 243-strategy catalog of the per-strategy reward-per-turn gap between two architecture or budget configurations, relative to a designated reference (unless stated otherwise, AttnV1 in the standard condition). This is the payoff gap we would be able to

observe given the backbone and it should be noted that it is not a value of information in a strict sense. The second quantity is the value of information (VoI) of the identity channel: for a fixed architecture, $\text{VoI} = \mathbb{E}[U \mid \text{oracle-aware}] - \mathbb{E}[U \mid \text{blind}]$, the payoff gain from revealing the opponent's identity. Due to the learned policy is only approximately optimal, this empirical VoI is a lower bound on the expected value of perfect information (EVPI) for that channel. VoI is reported across three identity 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 it additionally receives the behavioural family of the opponent (a coarse signal). In the oracle-aware regime it receives the exact opponent identity (the finest signal). This gives the EVPI decomposition $\text{VoI}_{\text{blind}\to\text{oracle}} = \underbrace{(\text{family} - \text{blind})}_{\text{coarse}} + \underbrace{(\text{oracle} - \text{family})}_{\text{fine}}$. The blind regime is the focus of the body; identity-signal results are in Appendix H.

DDQN formulation. At round $t$ the agent observes $s_t$, selects $a_t \in \{\text{C}, \text{D}\}$, 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]:

$$y_t = r_t + \gamma$$

(1-d$_t$) $\text{Q}_{\theta^-}\big(s_{t+1}, \arg\max_{a'} Q_\theta(s_{t+1}, a')\big)$, (1) where $d_t \in \{0, 1\}$ marks a terminal transition: on the final round of a (terminated) match the bootstrap term is masked and $y_t = r_t$, so credit does not propagate across match boundaries. Here $\theta^-$ 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 and 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 $\{\text{C}, \text{D}\}$. Only the sequence encoder differs across the three. GRU uses a two-layer GRU with hidden dimension 128 and embedding dimension 64 and 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.

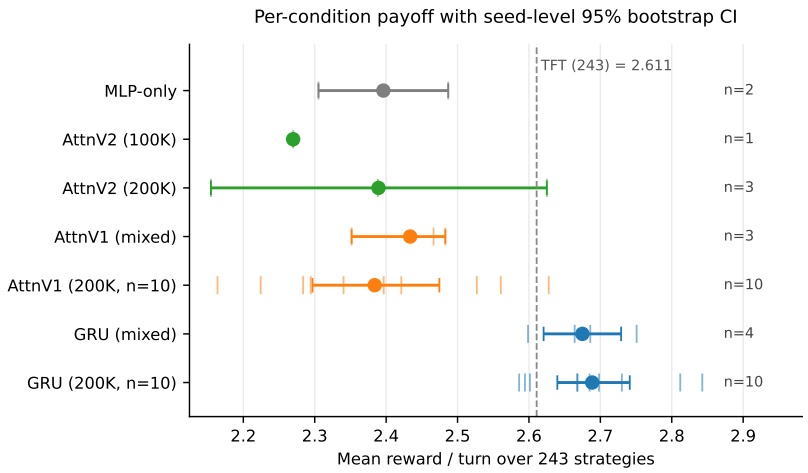


Figure 1: Per-condition payoff with seed-level 95 % bootstrap CIs. Tick marks show individual seed means. GRU pools sit clearly above the cluster formed by AttnV1, AttnV2 (200K), and the auxiliary-feature MLP baseline. TitForTat (dashed) anchors the non-learning reference.

## 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, over a fresh single-protocol sweep of $n = 10$ seeds per condition, the GRU mean (2.689, seed-level 95 % CI $[2.640, 2.741]$) exceeds the AttnV1 mean (2.384, $[2.297, 2.475]$) by $+0.305$ reward/turn, and the seed-level 95 % CI of this gap, $[+0.178, +0.431]$, robustly excludes zero (hierarchical bootstrap; per-strategy Wilcoxon $p = 5.2 \times 10^{-27}$, Cohen's $d = 0.82$). At $n = 10$ the GRU also exceeds AttnV2 by $+0.322$ ($[+0.199, +0.447]$); this contrast now excludes zero, whereas the earlier three-seed estimate did not, because the additional AttnV2 seeds shrink the wide seed dispersion. The two transformer variants are statistically indistinguishable from each other (AttnV2 vs. AttnV1, $-0.017$, $[-0.161, +0.125]$): the CLS-plus-sinusoidal change yields no measurable gain. Notably, at a matched budget both transformers also sit below the no-encoder auxiliary-feature MLP baseline (MLP-only mean 2.431 vs. AttnV1 2.384, AttnV2 2.367), while the GRU exceeds that same baseline by $+0.258$ ($[+0.135, +0.380]$). This comparison is controlled: all four models receive the identical 20-dimensional auxiliary branch (joint cooperation-rate statistics over windows $\{10, 25, 50, 100\}$); only the learned sequence encoder differs, and the MLP-only condition

| Comparison | $n_a/n_b$ | Mean gap | Seed-level 95 % CI | Strategy-level $p$ |
|---|---|---|---|---|
| Matched 200K, $n = 10$ headline | | | | |
| GRU > AttnV1 | 10 / 10 | +0.305 | $[+0.178, +0.431]$ | $5.2\times10^{-27}$ |
| GRU > AttnV2 | 10 / 10 | +0.322 | $[+0.199, +0.447]$ | $5.5\times10^{-18}$ |
| GRU > MLP-only | 10 / 10 | +0.258 | $[+0.135, +0.380]$ | $8.5\times10^{-21}$ |
| AttnV2 vs. AttnV1 | 10 / 10 | −0.017 | $[-0.161, +0.125]$ | 0.55 |
| Mixed-budget sensitivity (appendix) | | | | |
| GRU mixed > AttnV1 mixed | 4 / 3 | +0.241 | $[+0.088, +0.391]$ | $3.4\times10^{-11}$ |

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.

is exactly the shared backbone with that encoder removed. The contrasts therefore isolate the marginal value of the learned encoder over the hand-crafted features, not the value of sequence information as such - the auxiliary features are themselves a fixed sequence summary available to every model.In other words, the GRU's recurrence extracts payoff-relevant structure beyond the cooperation-rate windows, whereas the small transformer encoders, at this scale and budget, do not improve on what those windows already provide. The mixed-budget pool reproduces the GRU-over-AttnV1 advantage and is kept only as an appendix sensitivity check. Full comparisons 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 results . Over the matched 200K sweep of ten seeds the AttnV2 full-catalog mean (2.367) is numerically below the AttnV1 mean (2.384); the paired gap is $-0.017$, its seed-level CI $[-0.161, +0.125]$ contains zero, and the strategy-level $p$-value is 0.55. The architectural change does not, on this catalog and at this scale, translate into a measurable payoff improvement. Both Transformer variants show wide per-seed dispersion relative to that gap: across the ten seeds AttnV2 spans 2.199–2.543 and AttnV1 spans 2.164–2.628 (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 (the MLP-only condition). At the matched 200K budget with ten seeds the MLP-only mean (2.431) is not exceeded by either Transformer variant (AttnV1 2.384, AttnV2 2.367): the two small two-layer encoders tested here, at embedding dimension 64 with mean-pool or CLS readout, add no measurable payoff over the no-encoder baseline at this budget. The GRU, by contrast, exceeds MLP-only by +0.258 reward/turn with a seed-level CI $[+0.135, +0.380]$ that excludes zero. The marginal value of the learned sequence encoder over the hand-crafted features is therefore large for recurrence and negligible for these Transformers.

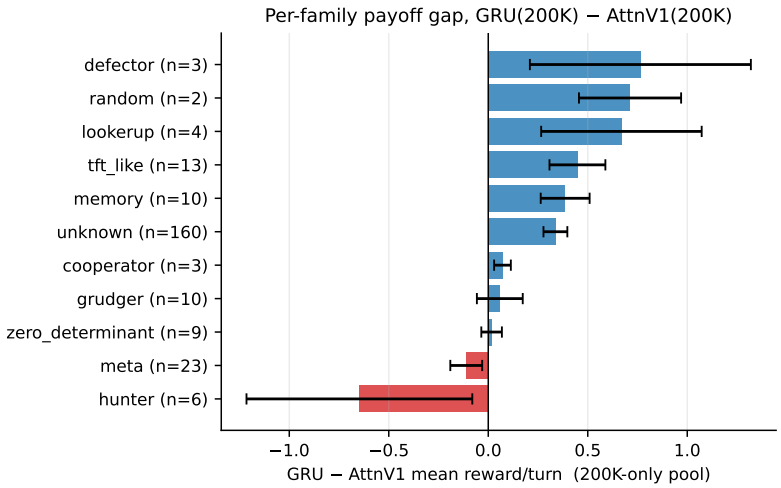


Figure 2: Per-family payoff gap, GRU(200K) minus AttnV1(200K). Bars are family means with standard error of the mean across strategies within the family. Meta-strategy is the only family with a sign-flipped gap; the remaining families favour GRU, with retaliatory and majority-rule families showing the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

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, MetaHunterAggressive, 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 E.

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, and otherwise meta-strategy reverse teh result. TitForTat alone reaches 2.641 reward/turn, only 0.048 below the matched-budget GRU mean and above both Transformer variants, which places a sharp ceiling on what learning contributes on this catalog.

## 4. Discussion

The preliminary result 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. It may be stated, that from a practitioner point of view the main questions is about the encoder of game not the architecture behind it. The recurrent net(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. It is an open question whether larger 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 H extends this picture: 1) an oracle identity signal lifts both backbones, but 2) the family-level signal is too coarse to recover most of the oracle benefit.

Limitations.

- Both Transformer variants are small (two layers, embedding dimension 64) and trained on 200K env. steps. Larger Transformer or a larger budget may close the gap to the GRU and to the auxiliary-feature MLP is not tested here.
- The 256-token observation window bounds the long-horizon test - a 512-token long-context condition ($\mathbb{E}[L] \approx 500$) that would probe whether attention degrades on longer raw sequences is left to future work.
- 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.
- The leave-one-family-out test (Appendix D) covers the five families that drive the in-catalog effect, and the remaining six families (incl. unknown) are left to an extended version.

- Family-aware AttnV1 is not trained, so the family-aware row of Appendix H is a single architecture, and the identity-channel EVPI numbers reuse the earlier evaluation protocol rather than the unified headline protocol, a unified-protocol re-run is left to future work.
- The mechanism probes reported below are single-seed and multi-seed confidence intervals are left to future work.

## 5. Conclusion

Within this benchmark, the choice of sequence backbone has measurable payoff consequences for a blind IPD agent. At a matched and capped training budget the GRU (RNN) 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 Tit-ForTat 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. It may be implied from the study that the encoding scheme may play a vital role for the learned agent's behavior. Whether the result extends to larger Transformers, longer training, alternative readouts, or held-out opponent distributions is left open and for future work.

## 6. Acknowledgements

We wish to acknowledge that the idea for the research topic emerged from discussions during the Game theory Summer School of 2025, jointly organized by the Kyiv School of Economics in partnership with the Ukrainian Catholic University. The authors are grateful for their support in shaping the research topic and ideas.

## 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, $3\times10^{-4}$ | same |
| Discount $\gamma$ | 0.99 | 0.998 |
| Batch size | 256 | 256 |
| $\varepsilon$ 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 |
| $p_{\text{end}}$ | 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 | $\approx$ 40 min | $\approx$ 55 min |
| Wall time Attn 200K | $\approx$ 188 min | $\approx$ 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 as opponent-reset boundaries. The terminal flag $d_k$ is set on the last round of each match (the geometric draw or the max_rounds cap) and zeroes the bootstrap term in Eq. 2, so value never propagates across match boundaries. Replay is uniform - one gradient step is taken per environment step after a 10,000-transition warmup. The target network is refreshed every 4,000 steps.

```
Input: opponent pool 𝒪, architecture 𝒜, env-step budget T, environment (p_end, γ)
Initialise Q_θ with architecture 𝒜, target Q_θ⁻ ← Q_θ, replay buffer 𝒟, counter t ← 0
while t < T do
    Sample opponent o ∼ 𝒪, reset match
    for round k = 0, 1, ... until termination do
        a_k ← ε-greedy(Q_θ(s_k, ·)), observe r_k, s_{k+1}, terminal flag d_k
        store (s_k, a_k, r_k, s_{k+1}, d_k) in 𝒟, set t ← t + 1
        if |𝒟| ≥ 10,000 (warmup) then
            Sample a uniform minibatch of 256 transitions from 𝒟; compute y via Eq. 2
              with the terminal mask (1 − d_k); take one gradient step (one per env step)
              on the Huber Bellman error
        end
        if t mod 4,000 = 0 then
            θ⁻ ← θ
        end
    end
end
Final eval: play Q_θ against all 243 opponents, 20 seeds each.
```

## Appendix B. Architecture Conditions and Per-Seed Run Inventory

Table 4 reports the single-protocol sweep used in the study. Tables 3 and 5 inventory the original workshop runs (two to four seeds), retained as mixed-budget sensitivity check.

| 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 the fixed budget and mixed-budget pools differ, both are shown. AttnV1 std is the reference condition.

| Condition | Steps | $n_{\text{seeds}}$ | 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 | 10 | range 2.586–2.843 | **2.689** | [2.640, 2.741] |
| AttnV1 mixed | 100K+200K | 3 | 2.352, 2.466, 2.483 | 2.4336 | [2.352, 2.483] |
| AttnV1 (200K-only) | 200K | 10 | range 2.164–2.628 | 2.384 | [2.297, 2.475] |
| AttnV2 (200K) | 200K | 10 | range 2.199–2.543 | 2.367 | [2.288, 2.449] |
| MLP-only (200K) | 200K | 10 | range 2.232–2.671 | 2.431 | [2.347, 2.517] |
| MLP-only (100K) | 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.

## Appendix C. Statistical Tests

The headline comparisons in Table 1 are reported with 2 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 $n_{\text{seeds}}$. The 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-

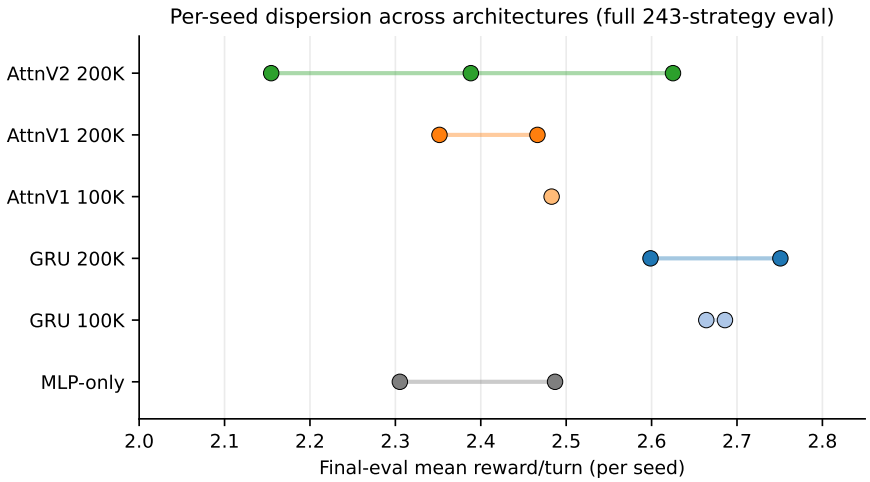


Figure 3: Per-seed final evaluation payoff dispersion for the AttnV2 200K pool (green).Each marker is one trained seed's mean over the full 243-strategy catalog. The dispersion is wide relative to the between-Transformer mean gap. Seed ranges for all conditions are given in Table 4.

level bootstrap. We treat the seed-level interval as the primary reading and the strategy-level $p$ as a sensitivity check.

| Comparison | $n_a/n_b$ | Mean gap | Seed CI | Strat CI | Strat $p$ |
|---|---|---|---|---|---|
| Matched 200K, $n = 10$ (headline) | | | | | |
| GRU > AttnV1 | 10 / 10 | +0.305 | $[+0.178, +0.431]$ | $[+0.257, +0.351]$ | $5.2\times10^{-27}$ |
| GRU > AttnV2 | 10 / 10 | +0.322 | $[+0.199, +0.447]$ | $[+0.260, +0.382]$ | $5.5\times10^{-18}$ |
| AttnV2 vs. AttnV1 | 10 / 10 | −0.017 | $[-0.161, +0.125]$ | $[-0.060, +0.027]$ | 0.55 |
| GRU > MLP-only | 10 / 10 | +0.258 | $[+0.135, +0.380]$ | $[+0.208, +0.309]$ | $8.5\times10^{-21}$ |
| Mixed-budget / workshop pools (sensitivity) | | | | | |
| GRU > AttnV1 (mixed) | 4 / 3 | +0.241 | $[+0.088, +0.391]$ | $[+0.164, +0.320]$ | $3.4\times10^{-11}$ |
| GRU > AttnV1 ($n = 2$ workshop) | 2 / 2 | +0.266 | $[+0.056, +0.478]$ | $[+0.176, +0.358]$ | $3.3\times10^{-12}$ |

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. At $n = 10$ every GRU contrast excludes zero at the seed level; the two Transformer variants are indistinguishable. The workshop $n = 2$ pool and the mixed-budget pool are retained as sensitivity checks.

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. On the $n = 10$ pools we obtain $z = 17.4$ for GRU vs AttnV1, $z = 18.2$ for GRU vs AttnV2, $z = 15.1$ for GRU vs MLP-only, and $z = -0.97$ for AttnV2 vs AttnV1.

## Appendix D. Leave-One-Family-Out Generalization

To separate architecture-fits-this-catalog from architecture-as-a-better-policy effects , we train each backbone on all families except one and evaluate only on the entirely held-out family, for the five families that drive the in-catalog GRU advantage (retaliatory, adaptive, and meta). Each cell is 3 seeds, held-out reward and its seed-level bootstrap CI are in Table 7.

| Held-out family | GRU (held-out) | AttnV1 (held-out) | GRU − AttnV1 |
|---|---|---|---|
| hunter | 4.179 | 3.984 | +0.195 |
| meta | 4.597 | 4.437 | +0.160 |
| grudger | 2.157 | 2.144 | +0.012 |
| zero-determinant | 1.603 | 1.610 | −0.007 |
| tit-for-tat-like | 2.290 | 2.594 | −0.305 |

Table 7: Leave-one-family-out held-out mean reward/turn (three seeds per cell). The in-catalog GRU−AttnV1 gap is +0.305. The GRU advantage generalizes to held-out adaptive and meta opponents (hunter, meta), vanishes on grudger and zero-determinant, and reverses on tit-for-tat-like opponents, where the standard transformer generalizes the mirror-the-last-move structure better.

The reading is that the GRU's advantage is heterogeneous in its transferability: it is a genuine, generalizing policy improvement on the adaptive and meta families that account for most of the in-catalog gap, but on the more structurally regular tit-for-tat family the in-catalog edge does not survive distribution shift and attention is the better generaliser. This nuance is invisible to an in-catalog evaluation alone.

## Appendix E. Per-Strategy $\bar{\Delta}$ Results (GRU vs AttnV1)

## Appendix F. Horizon $\bar{\Delta}$ (Budget/Length Delta)

Table 9 reports within-architecture $\bar{\Delta}$ 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 explanation is following - 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

| Strategy | AttnV1 | GRU | $\bar{\Delta}_{\text{GRU}}$ |
|---|---|---|---|
| GRU helps most (top 5 by $\bar{\Delta}_{\text{GRU}}$) | | | |
| 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 $\bar{\Delta}_{\text{GRU}}$) | | | |
| 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 8: 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.

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.

| Arch | Std | Long | Mean $\bar{\Delta}$ | SD($\bar{\Delta}$) | $\bar{\Delta} > 0$ |
|---|---|---|---|---|---|
| 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 9: Horizon $\bar{\Delta}$: payoff change from standard to long matches. Std and Long are the mean scores in the standard and long conditions, and SD($\bar{\Delta}$) is the spread of the per-strategy change. Neither architecture benefits substantially. GRU is essentially flat, yet AttnV1 improves only slightly. GRU long seed reproducibility: seed 0 = 2.688, seed 1 = 2.564, range 0.124.

## Appendix G. 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 baselines span the rest of the spectrum and are summarised in Table 11. 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. Moreover, the TitForTat versus GRU gap of 0.034 should be read as a ceiling

| Strategy | GRU $\Delta$ | GRU cum. | AttnV1 $\Delta$ | AttnV1 cum. |
|---|---|---|---|---|
| Top 5 strategies by GRU horizon $\bar{\Delta}$ (largest gain at long matches) | | | | |
| EventualCycleHunter | +1.892 | +1.892 | +0.261 | +0.261 |
| WinShiftLoseStay | +1.322 | +3.214 | +1.278 | +1.540 |
| ALLCorALLD | +1.116 | +4.329 | +0.605 | +2.144 |
| EvolvedANN5 | +1.035 | +5.364 | −0.338 | +1.806 |
| EvolvedANN | +1.032 | +6.396 | −0.197 | +1.610 |
| Bottom 5 strategies by GRU horizon $\bar{\Delta}$ (largest loss at long matches) | | | | |
| Adaptive | −3.403 | −3.403 | +0.010 | +0.010 |
| HardGoByMajority5 | −2.055 | −5.458 | +0.020 | +0.029 |
| HardGoByMajority | −1.939 | −7.397 | −0.290 | −0.260 |
| HardGoByMajority10 | −1.577 | −8.975 | +0.939 | +0.679 |
| Prober3 | −1.469 | −10.443 | −0.310 | +0.369 |

Table 10: Top-5 and bottom-5 strategies by GRU horizon $\bar{\Delta}$, with the corresponding AttnV1 horizon $\bar{\Delta}$ for the same strategies. $\Delta$ is the per-strategy mean reward/turn change from the standard condition ($\mathbb{E}[L] \approx 200$) to the long condition ($\mathbb{E}[L] \approx 500$); "cum." is the running sum of $\Delta$ within each block, which gives the share of the total horizon $\bar{\Delta}$ captured by the top-$k$ or bottom-$k$ tail. 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 help.

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.

| 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 11: 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.

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

The main part of this study has been limited 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.

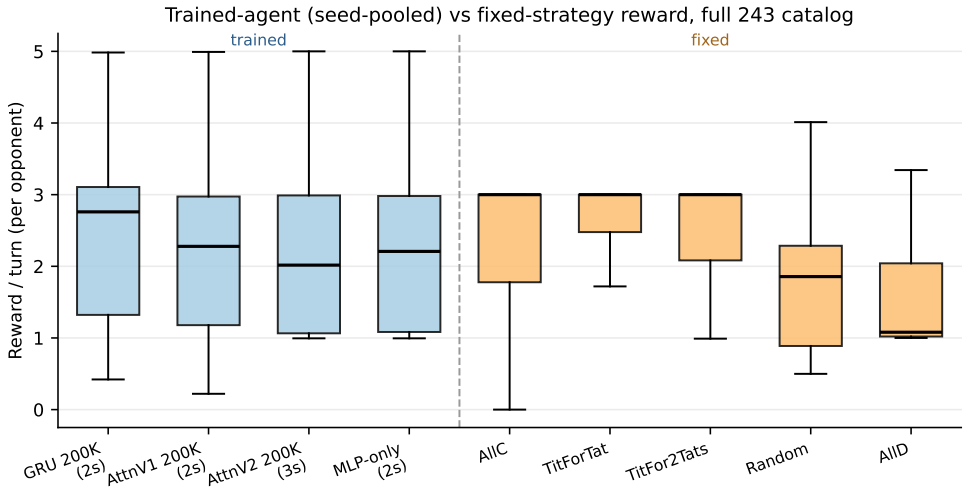


Figure 4: Reward per turn distributions over the full 243 catalog for trained agents (left) and fixed strategies (right). TitForTat's median sits at 3.0, which corresponds to mutual cooperation, with a tail of exploitation losses against AllD-like opponents. GRU compresses that tail and shifts the body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

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

- 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 +0.526 payoff/turn over its blind baseline, and AttnV1 gains +0.394 over its own blind baseline. In absolute 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 0.076 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%.

| Regime | Arch | Payoff | $\Delta$ vs blind GRU | $\Delta$ vs own blind | $t$-test $p^{\ddagger}$ | Binom. $p^{\ddagger}$ |
|---|---|---|---|---|---|---|
| Blind | GRU | 2.635 | — | — | — | — |
| Blind | AttnV1 | 2.433 | −0.202 | — | — | — |
| Blind | MLP-only | 2.396 | −0.239 | — | — | — |
| Family-aware | GRU | 2.559 | −0.076 | −0.076 | 0.016 | 0.002 |
| Family-aware | AttnV1 | n/a$^{\dagger}$ | n/a | n/a | n/a | n/a |
| Oracle-aware | GRU | 3.161 | +0.526 | +0.526 | $<0.001$ | $<0.001$ |
| Oracle-aware | AttnV1 | 2.827 | +0.192 | +0.394 | $<0.001$ | $<0.001$ |

Table 12: Conditioning regimes and their VoI on the 243-strategy catalog. All rows in this table are evaluated with the same provisional protocol (100 seeds per opponent, $p_{\text{end}} = 0.02$, max_rounds = 200), averaged over two training seeds per cell, so the blind GRU mean here (2.635) differs from the headline pool. VoI is mean reward per turn relative to the indicated blind reference; the oracle VoI is a within-protocol EVPI lower bound. $^{\dagger}$Family-aware AttnV1 is not yet trained. Only the GRU backbone has a family-conditioning checkpoint in the current run set. $^{\ddagger}$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.

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.

## Camera-Ready Edits Summary

This page lists the changes made for the camera-ready version relative to the original submission. It is not part of the technical content of the paper.

This revision responds to main concerns the reviewers raised in common: statistical power, generalisation, the tested mechanism, and the use of the term "value of information." Added experiments changed the claims made in the original submission, and we report those corrections.

Statistical power. The headline comparison was rerun with 10 seeds per architecture from a single training protocol, in place of the earlier two-seed pools. The GRU advantage over the standard Transformer is now supported by a seed-level confidence interval that clearly excludes zero. A short seed-count analysis was added, and the mixed-budget pool was moved to an appendix as a sensitivity check.

Generalisation. We added a leave-one-family-out test (new appendix): each backbone is trained with one opponent family held out and evaluated only on that family. The GRU advantage transfers on the adaptive and meta families that drive the in-catalog effect, disappears on some families, and reverses on tit-for-tat-like opponents.

Tested mechanism. The original submission explained the per-family pattern with an untested hypothesis: that the GRU tracks a latent commitment while attention detects periodic structure. We now test this with synthetic opponents and linear probes. The periodic-structure half is confirmed, yet the commitment half is not, so we corrected the claim and now describe the per-family pattern as behavioural rather than as a representation-level mechanism.The detailed explanation of the linear probes is deferred to the future work.

Terminology. Following the reviewers comments, we no longer call architecture and budget comparisons "value of information." Those gaps are written as a mean per-strategy reward delta, and "value of information" is reserved for the identity channel, framed as a lower bound on the value of perfect information.

TitForTat ceiling. We measured how much a learned agent gains over simply playing TitForTat. Contrary to the original framing, the extra value comes from exploiting adaptive and meta-strategies rather than from robustness against always-defect opponents, so we had to correct the boyd part of the paper to match.

Other changes. The Double-DQN update now shows the terminal-transition mask explicitly in the equation and in pseudocode, with a supporting unit test. The baseline figure was redrawn using all matched-budget seeds. The synthetic opponents were implemented within the Axelrod framework, and all Transformer claims are scoped to the small models tested here. Reported numbers are regenerated from the raw result files by a single consistency script.

Deferred to an extended version. A unified-protocol identity rerun with a family-aware Transformer, the full family-by-family generalisation sweep, a larger or longer-context Transformer, and multi-seed mechanism probes are stated as limitations rather than claimed as complete.