EXPLOITABILITY-GUIDED DISTILLATION OF INTERPRETABLE STRATEGIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual regret minimization computes strong strategies for imperfect-information games, but as large tables of mixed strategies that people can neither audit nor execute. Deploying them in human-facing settings means distilling them into small decision trees, and the standard objective for that distillation is a distributional distance to the teacher. We show this objective is misaligned with strategic quality: at the same size, a leaf fit by imitation can be far more exploitable than one fit by exploitability. We propose exploitability-guided distillation. The tree is grown best-first under an exploitability split score, and leaf strategies are optimized directly against per-player best responses, using the separability of NashConv into independent per-player objectives; a single scalarized objective then traces the safety-fidelity frontier. On Leduc poker, best-response leaf fitting is 2.3x to 4.3x less exploitable than imitation fitting across budgets of 4 to 32 leaves, and at $B = 8$ aligning the split score with the leaf objective cuts exploitability by 36 to 43 percent; the gain carries over to a five-rank variant. Our contributions are a bound that identifies the weights imitation misassigns to leaves, a distillation pipeline whose structure search and leaf solve share the exploitability objective, and an exact full-game evaluation showing that the leaf objective, with the teacher reduced to a uniform profile, returns the same Leduc structures at $B \leq 8$.

1 INTRODUCTION

Algorithms built on counterfactual regret minimization (CFR) (Zinkevich et al., 2007), including Monte Carlo variants (Lanctot et al., 2009) and discounted regret minimization (Brown & Sandholm, 2019a), have solved heads-up limit poker (Bowling et al., 2015) and beaten top professionals in far larger games (Moravčík et al., 2017; Brown & Sandholm, 2018; 2019b). Neural variants of CFR (Brown et al., 2019) and unified algorithms (Schmid et al., 2023) extend solving to still larger games. The strategies these systems produce are tables with one mixed action distribution per information set, often millions of entries. A human cannot audit such a table, a regulator cannot inspect it, and a training program cannot teach it. Whenever a strong strategy must be handed to people, it first has to be compressed into a small set of rules a person can read and simulate, in practice a decision tree with a few dozen leaves. Auditable reasoning is preferred in high-stakes settings (Rudin, 2019). Among the facets of interpretability (Lipton, 2018), the one we require of a distilled tree is exactly this.

The standard way to build that tree is imitation: fit a small student to the teacher's action distributions under a Kullback-Leibler (KL) or cross-entropy loss, as in knowledge distillation (Hinton et al., 2015), soft decision tree distillation (Frosst & Hinton, 2017), policy distillation (Rusu et al., 2016), subsequent tree-based policy extraction work (Bastani et al., 2018), and work that extracts human-understandable rules from computed poker strategies (Ganzfried & Yusuf, 2017). Interactive variants query the teacher online to avoid compounding errors (Ross et al., 2011). The implicit assumption is that a student close to the teacher in distribution is close in quality. In adversarial play that assumption fails. Exploitability, the payoff a best-responding opponent extracts, is a worst-case quantity. It is decided at the handful of information sets an adversary can attack, and those are exactly the ones a reach-weighted distance objective is happiest to ignore. Chasing distributional closeness under a tight budget can therefore preserve, or even concentrate, the student's strategic weaknesses. Restricting a strategy to a small, fixed class can itself introduce strategic pathologies (Waugh et al., 2009), and how to choose the restriction is open. Recent work learns it, from action abstractions chosen by

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

reinforcement learning (Li et al., 2024) to online pruning and abstraction of the game tree (Li & Huang, 2025).

This paper measures that misalignment and builds a pipeline that removes it where it is most damaging. We distill into mixed-strategy decision trees with a fixed leaf budget B: each leaf shares one behavioral mixed strategy across the information sets it covers, and splits test observable features. Leaf strategies are fit by minimizing, for each player separately, the value a best-responding (BR) opponent achieves; a one-line separability lemma, $\mathrm{NashConv} = \mathrm{BR}_0 + \mathrm{BR}_1$, makes this per-player decomposition exact. Structure is grown best-first, and the split score reuses a cheap version of the same leaf solve, so structure search and leaf fitting share one target. Whether that choice pays is an empirical question we answer in Section 5. Finally, a scalarized objective $\mathrm{BR} + \eta\,\mathrm{KL}$ sweeps the frontier between safety and faithfulness to the teacher.

The experiments separate the two choices (Figure 1). On Leduc poker with a near-Nash CFR+ teacher and 8 seeds, fitting leaves to the exploitability objective lowers NashConv by 2.3x to 4.3x over imitation fitting under both split scores at every budget from 4 to 32 leaves (interquartile mean, IQM). Distance-guided and exploitability-guided growth agree on the first split and diverge after it. Path overlap ends at most 0.11, and the information-set partitions they induce share 37 to 56 percent of their co-membership pairs, so the choice of split score decides which tree exists. Aligning the split score with the leaf objective cuts exploitability by a further 36 to 43 percent at $B = 8$, below the minimum of all 8 baseline seeds. A counterfactual ablation that swaps only the candidate prefilter back to an imitation score erases the entire gain. The gain carries over to the five-rank variant when every candidate is fine-scored: the aligned structure reaches 0.678 at $B = 8$, below all 8 KL-split seeds, whose minimum is 0.704. It disappears under the cheap two-stage variant with a top-K imitation prefilter (minimum 0.811, IQM 0.824), which is where an apparent reversal comes from. The safety-fidelity frontier is cheap up to a floor: at $\eta = 0.01$ the student captures 80 to 85 percent of the attainable fidelity gap under the reported 10^{-12} clip while NashConv rises by at most 0.016.

Our contributions are:

1. **We show that imitation is the wrong objective for leaf fitting.** We construct a two-state game where the imitation-optimal leaf is arbitrarily more exploitable than the exploitability-optimal leaf of the same size (Example 1), and we prove a one-sided bound (Proposition 1) whose weights are a best responder's reach and the local stake, not the teacher's.
2. **We make exploitability the target at every level of the pipeline.** We prove a separability lemma (Lemma 1) that reduces NashConv minimization to independent per-player leaf problems, and we build a derivative-free best-found leaf solver, a split score that reuses a cheap version of the same solve, and a scalarized safety-fidelity objective.
3. **We find that each alignment step pays.** A controlled study on Leduc, a five-rank variant with 780 information sets, and Kuhn poker: fitting leaves to the exploitability objective is consistently better on Leduc, by 2.3x to 4.3x at equal budget, and aligning the split score adds a further 36 to 43 percent at $B = 8$, carrying over to the five-rank variant when every candidate is fine-scored. A counterfactual ablation isolates the candidate prefilter as the cause of the one setting where it fails, and a teacher-free arm shows the teacher supplies only a warm start.

2 Related work

Abstraction and restricted strategy spaces. Poker research compressed games long before it compressed strategies. Card abstraction merges information sets before solving (Ganzfried & Sandholm, 2014; Gilpin & Sandholm, 2007; Kroer & Sandholm, 2014; Sandholm, 2015). Johanson et al. (2012) showed that within a fixed abstraction one should not settle for the abstract equilibrium: they compute the strategy in the restricted space that is best in the original game. We share their objective, the exploitability of a restricted strategy measured in the full game, but our restricted space is not given. We jointly select the tree structure and tests from a fixed set of observable features while fitting the leaf mixed strategies. The teacher supplies the initialization and the fidelity term; the objective itself reads the full game. Abstractions can mislead (Waugh et al., 2009), and evaluating them has its own literature (Johanson et al., 2013). We score candidate structures by exact full-game exploitability, so the split-score choice becomes an evaluated quantity, and our grouped CFR+ baseline tracks the

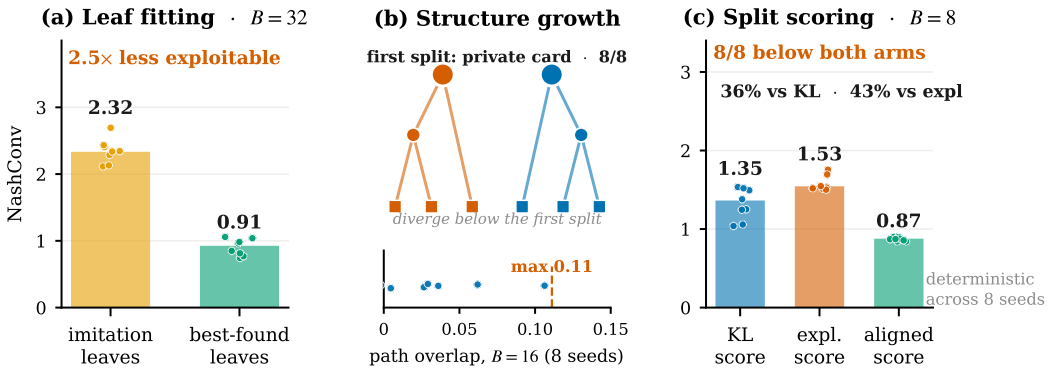

Figure 1: Imitation objectives misalign with exploitability at three levels of tree distillation on Leduc poker; bars are interquartile means over 8 seeds. (a) At a fixed exploitability-scored structure with $B = 32$, fitting leaves to the exploitability objective cuts NashConv from 2.32 to 0.91. (b) Growing with a KL score selects a different tree: both scores split on the private card first, then diverge. (c) Scoring splits with the best-found objective reaches NashConv 0.87 at $B = 8$, 36 percent below the KL arm and 43 percent below the exploitability arm.

imitation arms: solving the compressed game is not the same as minimizing full-game exploitability. A complementary line re-solves parts of the game, from safe nested subgame solving (Brown & Sandholm, 2017) to subgame solving without common knowledge (Zhang & Sandholm, 2021b). A second line learns the abstraction itself, by reinforcement learning (Li et al., 2024), online pruning of the game tree (Li & Huang, 2025), or a general information-abstraction framework (Li & Huang, 2026). These methods reduce or re-solve the game being solved. We keep the full game and the full teacher and reduce the strategy it represents.

Exploitability descent. Deep self-play learns strategy representations from scratch (Heinrich & Silver, 2016), and search-assisted deep RL reaches superhuman levels (Brown et al., 2020). Exploitability descent optimizes a strategy directly against best responses (Lockhart et al., 2019); learned best responses scale the idea to large games (Timbers et al., 2022). ApproxED maintains ensembles of candidate best responses during training (Martin & Sandholm, 2025), and constrained exploitability descent brings the objective to an offline reinforcement learning setting (Lu et al., 2025). These methods answer a solving question: drive a strategy without an interpretability budget toward equilibrium. We answer a compression question: given a strong teacher and a hard capacity limit, find the least exploitable student inside a structured class. The best-response machinery is shared; the object of study, the exploitability-versus-size trade-off of an interpretable class, is not.

Certificates. Nash certificates compact enough to check by hand exist for very large games (Zhang & Sandholm, 2020), and certified near-optimal strategies can be found in black-box games (Zhang & Sandholm, 2021a).

Distilling policies into trees. Tree distillation fits a small student to a teacher's outputs under a distance loss, as in behavioral cloning (Bain & Sammut, 1995), soft decision trees (Frosst & Hinton, 2017), and tree extraction for reinforcement learning policies (Bastani et al., 2018). Apprenticeship learning (Syed et al., 2008) also starts from expert behavior, matching the expert's feature expectations, where tree distillation matches action distributions. It likewise never evaluates a best response to the student. In card games, supervised learning from logs builds agents directly (Teófilo & Reis, 2011), and perfect-information distillation powers the strong DouDiZhu agent PerfectDou (Yang et al., 2022). In poker, Ganzfried & Yusuf (2017) extract human-understandable strategic rules from a computed strategy, and Bertsimas & Paskov (2022) fit supervised trees to CFR self-play data, with reported wins against the benchmark Slumbot. Rule languages capture such agents compactly (Teófilo et al., 2014; Lin & Hou, 2026; Hennes et al., 2026). Behavioral cloning, soft-tree distillation, and rule extraction treat faithfulness to the source strategy as the target and quality as a hoped-for consequence. Verifiable policy extraction weights fitting samples by Q values but still fits an imitation loss, so a best response to the student is not part of its objective (Bastani et al., 2018). What they

separate is the pair of objectives we do compare. On Leduc, imitation-fitted leaves on the same structures are 2.3x to 4.3x more exploitable than leaves fit to the exploitability objective, and the two objectives even disagree about which tree to grow. In adversarial domains the distillation loss is not a proxy for quality; it is a different objective with a different optimum. Markov-game theory reaches the same conclusion: offline cloning can miss the equilibrium (Freihaut et al., 2025), and zero-error imitation can stay arbitrarily exploitable (Bergerault et al., 2026). We localize the failure to leaf fitting inside an EFG tree. Further background is in Appendix A.

3 PRELIMINARIES

We work in finite two-player zero-sum extensive-form games (EFGs) with perfect recall. An EFG has players $p \in \{0, 1\}$, chance nodes, and for each player a partition of decision nodes into information sets $\mathcal{I}_p$. The acting player observes only the information set, not the node. A behavioral strategy σ_p assigns each $I \in \mathcal{I}_p$ a distribution over its actions, and $\sigma = (\sigma_0, \sigma_1)$ is a profile. Reaching an information set has a probability we call its *reach*. The player's own reach $\pi_p^\sigma(I)$ multiplies only p's action probabilities on the path to I; the chance-and-opponent reach $\pi_{-p}^\sigma(I)$ multiplies the chance probabilities and the opponent's. A reach-weighted average weights each information set by $\pi_{-p}^\sigma(I)$. The exploitability of a profile, written $\mathrm{NashConv}$ and reported below as NashConv, is

$$\mathrm{NashConv}(\sigma) \;=\; \sum_{p\in\{0,1\}} \Big[\max_{\sigma'_p} u_p(\sigma'_p, \sigma_{1-p}) \;-\; v_p^{\mathrm{NE}} \Big] \;=\; \mathrm{BR}_0(\sigma_1) + \mathrm{BR}_1(\sigma_0), \tag{1}$$

where $\mathrm{BR}_p(\sigma_{1-p})$ is the best-response value net of the equilibrium value v_p^{NE}, so that $\mathrm{NashConv} = 0$ exactly at a Nash equilibrium. Two conventions appear below and we keep them apart. The raw best-response payoff u_p^{BR} differs from BR_p by the game value and is the quantity quoted for individual players ($13/30$ in Appendix D); BR_p is the shifted value used in every NashConv.

A teacher $\sigma^\star$ is a near-equilibrium profile produced by a standard solver; we use CFR+ (Tammelin et al., 2015). The student class is the set of mixed-strategy decision trees with leaf budget B, written $\mathcal{T}(B)$. A tree in $\mathcal{T}(B)$ routes every information set of a player through binary tests on observable features of that information set (private card, board card, betting round, raise count, action faced) to one of at most B leaves. Each leaf holds a single behavioral mixed strategy shared by all information sets routed to it. Leaves only group information sets with the same number of actions, and the shared distribution is read by position. In Leduc the two-action sets are check/raise when no bet is faced and fold/call at the raise cap (the per-round limit on raises), so the first entry means put no chips in and the second means put chips in. Without this restriction a leaf's distribution would be undefined at some information sets; it shapes the initial tree (one leaf per player and action count) and holds by construction throughout growth. Each rule then reads as one executable instruction: in every situation matching these conditions, play this action distribution. One can simulate the student by walking at most $B - 1$ tests (far fewer on the balanced trees growth produces) and mixing as instructed.

The problem is

$$\min_{\sigma\in\mathcal{T}(B)} \mathrm{NashConv}(\sigma), \tag{2}$$

with $\sigma^\star$ available as guidance. Imitation distillation replaces (2) by minimizing a reach-weighted distance to $\sigma^\star$, typically the forward KL $\mathrm{KL}(\sigma^\star \,\|\, \sigma)$, whose closed-form optimum at fixed structure is the reach-weighted average of the teacher inside each leaf. Every KL below floors both of its inputs at 10^{-12} and renormalizes before the sum, so an action the teacher plays with probability zero still yields a finite distance; the floor moves that average by an amount of order 10^{-12}. The next sections quantify how far that substitution is from the actual objective.

4 EXPLOITABILITY-GUIDED DISTILLATION

The pipeline has three stages: grow the structure, fit the leaves, and optionally trade safety against fidelity. The leaf objective is the quantity we are judged on, NashConv in the original game minimized per player. The split score reuses a cheap approximation of the same solve, and Section 5 tests whether that choice helps.

Lemma 1. *For any profile $\sigma = (\sigma_0, \sigma_1)$ in a two-player zero-sum EFG,* $\mathrm{NashConv}(\sigma) = \mathrm{BR}_0(\sigma_1) + \mathrm{BR}_1(\sigma_0)$*, and the two terms depend on disjoint parts of the profile.*

Proof. The identity is Equation (1). For the second claim, the best-response value $\mathrm{BR}_p(\sigma_{1-p})$ is computed by a backward walk that maximizes over player p's actions at p's information sets and takes expectations under σ_{1-p} and chance elsewhere; the walk never reads σ_p. Hence BR_1 is a function of σ_0 alone and BR_0 of σ_1 alone. □

At a fixed structure, the leaf strategies of the two players can therefore be optimized independently. One coupled minimization over all leaves becomes two problems of half the size, which makes a direct attack on (2) tractable (numerical check in Appendix E).

Growth starts from the coarsest legal tree, one leaf per player and action count, and repeatedly applies the best-scoring split anywhere in the current tree, so the budget counts total leaves, at any depth. Each candidate splits one leaf's information sets by a feature test; the intermediate trees at every leaf count $4, 5, \ldots, B$ form one incremental path, so a single run yields the whole exploitability-versus-budget curve. For seed diversification the experiments select uniformly among candidates within one percent of the best score, using a per-seed random stream.

The baseline scores are the lifted score, which lifts the imitation-fitted leaves to a full-game profile and takes the exact NashConv of the tentative tree (labeled expl. in figures), and the KL score, the symmetric KL to the teacher, summed in both directions and averaged over the teacher's information sets without reach weights. Both are cheap because the imitation fit is closed form. The KL baseline therefore pairs an unweighted split score with a reach-weighted leaf fit; a reach-weighted forward KL, the score the imitation leaf fit itself minimizes, is no better at any budget and worse from $B = 8$ (Table 11). The lifted score is exploitability-aware but evaluates candidates under a leaf fit that the final solve will replace; this detail is where Section 4 intervenes. We refer to the growth procedure with the one-percent tolerance and per-seed stream as the run protocol.

For a fixed structure, Lemma 1 reduces the problem to, for each player p,

$$\min_{x_p} \; f_p(x_p) \;=\; \mathrm{BR}_{1-p}\big(\sigma_p(x_p)\big), \tag{3}$$

where x_p stacks the leaf distributions of player p and $\sigma_p(x_p)$ is the induced behavioral strategy. The objective reads the teacher nowhere: $\sigma^\star$ enters only through the initialization and through the fidelity term of Section 4. We minimize (3) with derivative-free Powell search from five starts (the imitation fit, a CFR solve of the restricted game, the uniform strategy, and two fixed-seed Dirichlet draws) and return the best. We call the solver best-found: the objective is piecewise smooth, and multi-start local search carries no global optimality guarantee.

The exploitability-aware split score of Section 4 still evaluates candidates with imitation-fitted leaves, while the final leaves are fit by (3). The two stages judge structures by different criteria, and Section 5 shows they disagree sharply. The aligned score judges candidates by (3) itself; it grows trees whose root-to-leaf paths overlap the lifted-grown and KL-grown trees in at most 0.44 and 0.27 of the measured path sets at $B = 8$, and at most 0.07 at $B = 16$ over both baselines (Appendix E). The aligned score removes the mismatch by evaluating every candidate split with a cheap best-found solve of (3) itself, one start from the imitation fit and a small iteration cap. For the cost analysis we also define a two-stage variant: the lifted score ranks all candidates as a prefilter, and only the top $K = 8$ reach the aligned fine score. It cuts growth cost by roughly the candidate count over K and reintroduces the mismatch at the ranking stage, which Section 5 turns into a counterfactual test.

A deployed student may need to stay recognizably close to the teacher. We scalarize the two goals per player,

$$f_p^\eta(x_p) \;=\; \mathrm{BR}_{1-p}\big(\sigma_p(x_p)\big) \;+\; \eta\, F_p(x_p), \tag{4}$$

where the fidelity gap F_p is the reach-normalized forward KL from the teacher to the student on player p's information sets, so smaller is closer to the teacher. Setting $\eta = 0$ recovers (3) exactly; sweeping η traces the trade-off between NashConv and fidelity. The identity $f_0^\eta + f_1^\eta = \mathrm{NashConv} + \eta(F_0 + F_1)$ serves as an implementation check (Appendix E).

Why can imitation fitting still leave a student exploitable? A two-state example isolates the mechanism before Proposition 1 generalizes it.

Example 1. Chance selects state A with probability $1 - \epsilon$ and state B with probability ϵ; player 1 acts in both, and the two states are distinct information sets with the same action set $\{a_1, a_2\}$. In A, action a_1 ends the game with payoff $\delta > 0$ to player 1 and a_2 ends it with payoff 0. In B,

after a_1 the opponent can punish, giving player 1 payoff $-M < 0$, while a_2 again yields 0. The teacher plays a_1 in A and a_2 in B. Distill player 1 into a single shared leaf that must play one distribution $(q, 1-q)$ over (a_1, a_2) in both states. The imitation-optimal leaf is the reach-weighted teacher average, $q = 1 - \epsilon$. The punishing branch in B is then reached with probability $\epsilon q > 0$, and the best-responding opponent punishes there, giving exploitability that grows linearly in M. The exploitability-optimal leaf plays $q = 0$ once $\epsilon M > (1-\epsilon)\delta$, with exploitability $(1-\epsilon)\delta$ independent of M, and the ratio $\epsilon(\delta + M)/\delta$ is unbounded in M (derivation in Appendix B).

The example isolates a general weight mismatch. Write TV for total variation; with the notation of Section 3, $\pi^{\mu}_{-p}(I)$ is the chance-and-opponent reach of I under μ, summed over its histories.

Proposition 1. *Let σ and τ be profiles and let μ_p be a best response of player $1-p$ to σ_p. For $I \in \mathcal{I}_p$ and $a \in A(I)$, let $q_I(a)$ be player $(1-p)$'s counterfactual value of action a at I when p plays σ_p above I and τ_p below it, normalized by $\pi^{\mu_p}_{-p}(I)$; a set with $\pi^{\mu_p}_{-p}(I) = 0$ contributes zero. Let $r_I \le \Delta$ be the range of q_I over a, with Δ the utility range. Then*

$$\mathrm{NashConv}(\sigma) - \mathrm{NashConv}(\tau) \;\le\; \sum_{p} \sum_{I \in \mathcal{I}_p} \pi^{\sigma}_{p}(I)\, \pi^{\mu_p}_{-p}(I)\, r_I\, \mathrm{TV}\big(\sigma_p(I), \tau_p(I)\big). \tag{5}$$

The proof (Appendix C) fixes μ_p, telescopes the value difference over p's information sets, and bounds each local term by r_I times the total variation. Imitation weights a per-set divergence by $\pi^{\tau}_{-p}(I)$, the reach of the teacher's opponent, and by nothing else. The bound weights the per-set distance by the student's own reach, by the reach of a best responder to the student, and by the local stake r_I, which in Example 1 is $r_B = M$. On the aligned $B = 8$ Leduc structure the two reach weightings, with r_I set aside, differ by total variation 0.23 to 0.34, and 13 to 37 percent of the bound's reach weight falls on information sets that carry under one percent of the imitation weight (Appendix C).

5 Experiments

The main testbed is Leduc poker (Southey et al., 2005): 288 information sets in our encoding, private and public cards, two betting rounds, and raise caps. Kuhn poker (Kuhn, 1950) is a sanity check. A five-rank variant, Leduc(5,2), probes scale: ten cards, 780 information sets, 2.7 times the original game, the same betting structure and seeds. The teacher is CFR+ run for 3000 iterations, with NashConv 1.4×10^{-4} (6.8×10^{-4} on the cluster runs listed in Appendix L), at least three orders of magnitude below any student reported here. All arms share the same 8 seeds, which randomize the structure search through the one-percent score tolerance of Section 4 against a teacher fixed per run. Reported values are exact NashConv from full best-response traversal on the original game; aggregates are interquartile means (IQM) with interquartile bands. The final leaf solve uses the identical configuration in every experiment, so numbers are comparable across tables; the full parameter table and the three growth protocols are in Appendix L.

Cost on one CPU core: an imitation-scored candidate evaluates in 10 to 20 milliseconds and an aligned-scored candidate in about 20 seconds, over 64 to 128 candidates per step on Leduc. The aligned score takes 74 minutes per seed to $B = 8$ and 4.5 to 10 hours to $B = 16$. On the five-rank variant the $K = 8$ prefiltered arm reaches $B = 16$ in 6 to 9 hours, while the lifted and KL scores keep a scan seed under 90 minutes. Scoring every candidate takes 36 to 39 minutes per seed on 15 workers at $B = 8$ (Appendix H).

Leaf-level misalignment. Figure 2 and Table 1 report the factorial comparison: two split scores (exploitability-aware and KL) crossed with two leaf solves (imitation and best-found), plus a grouped CFR+ heuristic (Appendix L) on the exploitability-scored structure. Both best-found arms dominate both imitation arms at every budget; at $B = 32$ the gap is a factor of 2.5 under the exploitability split score and 3.9 under the KL score (the ratio rows of Table 1). The gap is not the weighting: reweighting the same fit by the teacher's own advantage makes every budget *more* exploitable (Appendix J). Liar's Dice, a game outside the Leduc family, reproduces the ordering on every seed at every budget (Appendix K).

The restricted-game baseline, grouped CFR+ on the same leaves, lands with the imitation arms. It is a heuristic solver: leaf sharing can merge information sets that differ in their own past actions, so

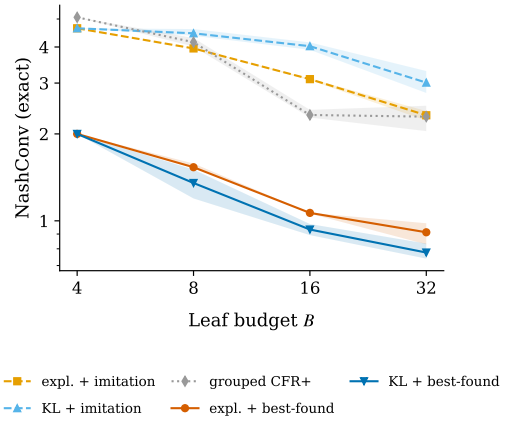

Figure 2: Exact NashConv against leaf budget on Leduc (IQM line, interquartile band, 8 seeds per arm). Both best-found arms dominate both imitation arms and the grouped CFR+ baseline at every budget.

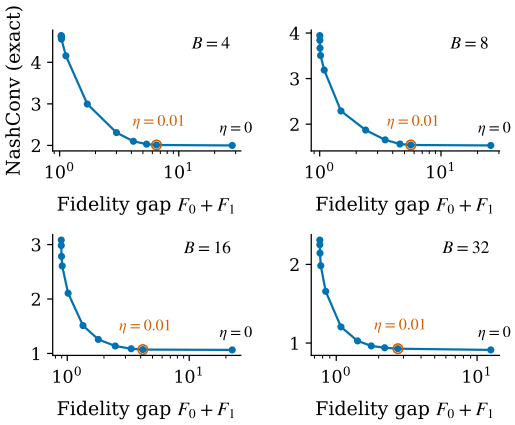

Figure 3: Safety-fidelity frontier of the scalarized objective (IQM over 8 seeds, one panel per budget). From $\eta = 0$ to $\eta = 0.01$ the student captures 80 to 85 percent of the attainable gap under the reported 10^{-12} clip, while NashConv rises at most 0.016. Vertical scales differ per panel.

Table 1: Exact NashConv of the distilled tree on Leduc (IQM over 8 seeds). Best-found leaf fitting beats imitation at every budget under both split scores; the ratio rows give the factor (smaller is better). The grouped CFR+ baseline trails both best-found arms.

Arm	$B = 4$	$B = 8$	$B = 16$	$B = 32$
expl. split + imitation leaves	4.644	3.954	3.096	2.321
KL split + imitation leaves	4.644	4.461	4.026	3.011
grouped CFR+ (restricted game)	5.068	4.156	2.326	2.295
expl. split + best-found leaves	2.000	1.533	1.065	0.913
KL split + best-found leaves	2.000	1.352	0.933	0.777
ratio imitation/best-found (expl. split)	2.32x	2.58x	2.91x	2.54x
ratio imitation/best-found (KL split)	2.32x	3.30x	4.32x	3.88x

perfect recall need not survive the grouping (Appendix L). Its objective is the compressed game, not the full game we evaluate, a gap Johanson et al. (2012) identify for fixed abstractions. We adopt their objective and read its shortfall as objective difference mixed with its own optimization quality. Kuhn poker shows the same gap on the handcrafted six-leaf grouping: grouped CFR+ yields NashConv 0.364, minimizing NashConv over the same leaves yields 0.138, and learning the structure improves this further to 0.056 at equal budget (protocol and teacher in Appendix F).

Two features of the ratios point at the leaf solve as the source. Under the KL split score the imitation gap is wider at every budget above 4 (3.9 versus 2.5 at $B = 32$): when the structure is grown by a distance objective, the leaves a reach-weighted fit is happiest to ignore are the ones a best responder attacks. The ratio is also not monotone in the budget, though every arm's NashConv falls steadily.

Structure-level misalignment. If the two split scores merely ranked the same structures differently, leaf fitting could compensate for the choice; they do not rank the same structures. Re-growing the lifted and KL trees at $B = 16$ for all 8 seeds and comparing them with the metrics of Appendix E shows the divergence is not a single root disagreement. Both scores make the same first split, on the private card, in all 8 seeds, and accumulate step by step. Whole-tree path Jaccard, a syntactic measure that scores each root-to-leaf condition sequence, ends between 0 and 0.11, exactly zero in 3 of 8 seeds (median 0.03). The partitions the trees induce over information sets are less disjoint: the two scores share 37 to 56 percent of their co-membership pairs at $B = 16$ and 46 to 68 percent at $B = 8$, with Rand indices of 0.85 to 0.92 throughout. Most of that Rand agreement is agreement that

pairs belong to *different* leaves, so the co-membership Jaccard is the sharper reading: the two scores group in common at most 0.68 of the pairs they place at all at $B = 8$ and 0.56 at $B = 16$. The leaf solve sees the grouping, not the tests that express it. The two scores also rely on different feature families: KL-grown trees split almost exclusively on card features, and exploitability-grown trees mix in action-history features. The simpler family is also the better one under the full best-found leaf solve at $B \geq 8$ (Table 1), though not uniformly per seed.

Aligning the split score with the leaf solve. Table 2 evaluates the aligned split score of Section 4. At $B \leq 8$ the tolerance never binds and all seeds grow one deterministic tree (Appendix G), so the aligned structure is compared against the 8 distinct trees of each baseline arm. It falls below all 8 outcomes of both baseline arms, 36 to 43 percent below the baseline IQMs. The gain is not from determinism: giving the baselines the identical single-tree protocol (zero tolerance, deterministic argmin) keeps them above the aligned structure (Table 2). At $B = 16$ the tolerance binds, and all 8 aligned seeds (0.763 to 0.841) lie below all 8 KL-score seeds, lowest 0.858 (Table 5).

The counterfactual row isolates the cause. Keeping the aligned fine score and the identical leaf solve but letting the lifted score rank all candidates first, with only the top K reaching the fine score, erases the entire gain. At $K \in \{2, 4, 8\}$ every seed lands on NashConv 1.533, the value of the zero-tolerance lifted-score tree (Table 6). At $K = 16$ the prefilter starts to leak the gain, with one seed at 1.067 and the rest between 1.478 and 1.533. Once the prefilter controls the shortlist, the fine score can only choose among candidates the mismatched ranking already favored: the benefit lives in the ranking of candidate splits. The five-rank variant below runs precisely this prefiltered arm, and the reversal it shows there at small budgets disappears when every candidate is fine-scored.

A compute-matched control shows this restart allocation does not close the gap. Granting the KL score the aligned arm's CPU budget, spent on 153 to 269 tolerance-randomized KL growths ranked by a cheap solve with only the winner fully solved, leaves all 8 seeds at NashConv 0.967, above the aligned arm's 0.867. At matched CPU the growth count does not separate the arms; the ranking does.

Removing the teacher end to end does not change the outcome on Leduc. With the uniform profile in every teacher role and no teacher solve, the same protocol returns the identical structure at $B \leq 8$ (all 8 seeds, to machine precision) and IQM 0.792 against 0.807 at $B = 16$. On the five-rank variant the teacher-free arm matches the all-candidate value on 4 of 8 seeds and lands 0.03 above it on the rest (IQM 0.693 against 0.678). The teacher supplies a warm start (Appendix G).

Appendix G leaves the aligned result unchanged at $B \leq 8$ under teacher strength 1000 to 6000 iterations and score tolerance 1 to 10 percent.

The aligned structure is readable in full: Table 3 lists its four rules per player, each reading only bet faced, private card, and board dealt. Every leaf plays one action with probability at least 0.97, and rounding all leaves to their most likely action leaves NashConv unchanged at 0.867 (Appendix D). The student executes as a pure list of eight rules.

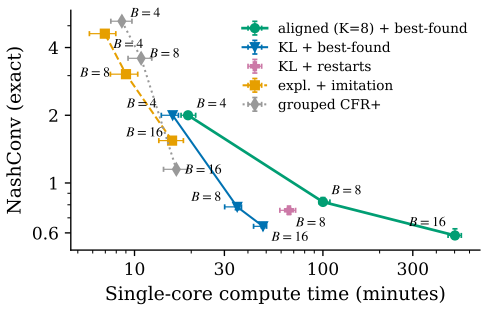

Figure 4: Quality against single-core time on the five-rank variant (IQM over 8 seeds, B labeled, both axes logarithmic). Both best-found arms dominate the imitation and grouped CFR+ arms; their order flips with budget. The aligned arm is the $K = 8$ prefiltered one.

Figure 4 reports the same four arms on the larger variant as a quality-versus-time curve. The two best-found arms dominate the imitation arm at every budget, and their order flips. At $B = 8$ the KL arm is lower on 6 of 8 seeds and tied on one; at $B = 16$ the aligned arm is lower on 7 of 8, at tenfold cost in single-core time. The aligned arm here is the prefiltered one, because fine-scoring every candidate at every step costs far more on this game. Scoring every candidate removes the small-budget reversal: the aligned structure then reaches NashConv 0.678 at $B = 8$, below every KL seed, whose minimum is 0.704 (Table 7). The reversal is a property of the candidate prefilter, not of alignment. The four growth steps produce one or two distinct partitions each across the 8 seeds, six in all. All 8 seeds converge on the same 8-leaf partition, so the 32 replayed seed-steps are not 32 independent states. The committed candidate ranks 2 to 20 of 96, 3 to 14 of 106, 74 to 90 of 116,

Table 2: Aligning the split score with the leaf-solve objective on Leduc ($B = 8$, full best-found leaf solve). The aligned arm grows one deterministic structure at $B \leq 8$, so its NashConv is compared against all 8 baseline trees per arm. The last two columns count, out of the eight trees of each baseline arm, how many this row's value falls below; a dash marks a self-comparison.

Split score	IQM NashConv	beats KL	beats expl.
aligned split (pure)	0.867	8/8	8/8
KL split, zero tolerance	1.230	—	8/8
KL split	1.352	—	8/8
expl. split, zero tolerance	1.533	0/8	—
expl. split	1.533	0/8	—
aligned + lifted prefilter	1.533	0/8	2/8

Table 3: The aligned $B = 8$ tree on Leduc, the deterministic structure of Table 2. Both players grow the same four rules over whether a bet is faced, the private card (2 is the top rank), and whether the board is dealt; the rules are first-match in the order listed, so the first row absorbs the raise-cap information sets; probabilities are player 0 / player 1 and size counts each player's information sets. Rounding every leaf to its most likely action leaves NashConv unchanged at 0.867.

Situation	Conditions	Action	Probability	Size
no bet faced, or raise cap reached	none	check; fold at the cap	1.00 / 1.00	96
facing a bet	$card \neq 2$	fold	1.00 / 1.00	32
facing a bet	$card = 2$, board dealt	call	1.00 / 1.00	15
facing a bet	$card = 2$, no board yet	raise	0.99 / 0.97	1

and 73 to 90 of 113 legal candidates under the lifted score (median 46.5), part of the spread being equivalent encodings of one partition (a card test and its complement). A top-8 prefilter keeps that exact candidate on 2 seeds at the first step, 6 at the second, and none afterwards, though a shortlist can still contain a different encoding of the same partition. The compute-matched control agrees. Under that arm's per-seed budget, 10 to 16 randomized KL growths and one full solve reach IQM 0.756 within 78 percent of it, below the single KL pass at 0.783 and the prefiltered arm at 0.824.

The safety-fidelity frontier. Figure 3 sweeps η over 11 values at four budgets. The sampled points are monotone at every budget: across all 320 consecutive grid pairs, NashConv never decreases and the fidelity gap never increases. Each structure also has a closed-form fidelity floor $F_{\min}$, the reach-weighted teacher average that minimizes the reach-weighted forward KL on a fixed grouping. The sweep reads most sharply as the share of the gap $F(0) - F_{\min}$ it takes: at $\eta = 0.01$ the student captures 80 to 85 percent of that gap (69 to 87 percent per seed) while NashConv rises by at most 0.016, the unregularized solve sitting 11.7 to 27.1 above a floor that is itself 0.76 to 1.03. That share is read under the reported 10^{-12} clip. At a 10^{-4} floor it shrinks to 31 to 52 percent, while the budget IQMs of the floor and the regularized endpoint move by at most a percent, the unregularized excess living in probabilities below 10^{-4} (Appendix L). The expensive region is the far end, where the sweep reaches the floor to within 1.3×10^{-5} and exploitability has risen several-fold. The floor belongs to the grouping; Table 12 gives it for the four budgets (eight seed structures each) and the payoff scale tested here. The aligned structures show the same cheap first step (Appendix L).

6 Conclusion

Distilling a strong strategy into a readable tree is usually framed as imitation. For leaf fitting that framing is wrong at every Leduc budget we tested, costing a factor of 2.3 to 4.3 in exploitability, and Proposition 1 bounds the gap with weights unlike those imitation assigns. Scoring candidate splits by the objective of the final leaf solve, with fidelity as a cheap regularizer, returns one deterministic eight-rule structure (Table 3) below all 8 baseline-arm outcomes at $B = 8$, and the gain survives removing the teacher. At $\eta = 0.01$ the student captures 80 to 85 percent of the gap to its closed-form floor for a NashConv rise of at most 0.016.

AI USE STATEMENT

Generative AI tools assisted with language polishing, with writing and debugging parts of the experiment code, and with running experiments and checking their outputs, including the scripts that produce the figures and tables. The research design, the mathematical derivations, and the interpretation of the results are the authors' own, and we take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

An anonymized code copy accompanies this submission: the game engine, the CFR+ teacher, exact best-response evaluation, the drivers for Section 5 and Appendix K, and the baselines they are compared against. Drivers ship as compiled Linux x86-64 extensions behind a short readable launcher, so the numerical core is not shipped as source. NashConv values are exact full-traversal best-response evaluations, and hyperparameters and per-seed protocols are in Appendix L.

REFERENCES

Michael Bain and Claude Sammut. A framework for behavioural cloning. In *Machine intelligence 15*, pp. 103–129, 1995.

Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. Verifiable reinforcement learning via policy extraction. *Advances in neural information processing systems*, 31, 2018.

Antoine Bergerault, Volkan Cevher, and Negar Mehr. Matching multiple experts: On the exploitability of multi-agent imitation learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=B9qYvPTbTN.

Dimitris Bertsimas and Alex Paskov. World-class interpretable poker. *Machine Learning*, 111(8): 3063–3083, 2022.

Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin. Heads-up limit hold'em poker is solved. *Science*, 347(6218):145–149, 2015.

Noam Brown and Tuomas Sandholm. Safe and nested subgame solving for imperfect-information games. *Advances in neural information processing systems*, 30, 2017.

Noam Brown and Tuomas Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

Noam Brown and Tuomas Sandholm. Solving imperfect-information games via discounted regret minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1829–1836, 2019a.

Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456): 885–890, 2019b.

Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep counterfactual regret minimization. In *International conference on machine learning*, pp. 793–802. PMLR, 2019.

Noam Brown, Anton Bakhtin, Adam Lerer, and Qucheng Gong. Combining deep reinforcement learning and search for imperfect-information games. *Advances in neural information processing systems*, 33:17057–17069, 2020.

Till Freihaut, Luca Viano, Volkan Cevher, Matthieu Geist, and Giorgia Ramponi. Learning equilibria from data: Provably efficient multi-agent imitation learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=QNtJO0jTf3.

Nicholas Frosst and Geoffrey E. Hinton. Distilling a neural network into a soft decision tree. In *CEx@AI*IA*, volume 2071 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2017.

Sam Ganzfried and Tuomas Sandholm. Potential-aware imperfect-recall abstraction with earth mover's distance in imperfect-information games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

Sam Ganzfried and Farzana Yusuf. Computing human-understandable strategies: deducing fundamental rules of poker strategy. *Games*, 8(4):49, 2017.

Andrew Gilpin and Tuomas Sandholm. Lossless abstraction of imperfect information games. *Journal of the ACM (JACM)*, 54(5):25–es, 2007.

Johannes Heinrich and David Silver. Deep reinforcement learning from self-play in imperfect-information games. *arXiv preprint arXiv:1603.01121*, 2016.

Daniel Hennes, Zun Li, John Schultz, and Marc Lanctot. Code-space response oracles: Generating interpretable multi-agent policies with large language models. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '26, pp. 3525–3527, Richland, SC, 2026. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400723179. doi: 10.65109/IKJF6607. URL https://doi.org/10.65109/IKJF6607.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Michael Johanson, Nolan Bard, Neil Burch, and Michael Bowling. Finding optimal abstract strategies in extensive-form games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pp. 1371–1379, 2012.

Michael Johanson, Neil Burch, Richard Valenzano, and Michael Bowling. Evaluating state-space abstractions in extensive-form games. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pp. 271–278, 2013.

Christian Kroer and Tuomas Sandholm. Extensive-form game abstraction with bounds. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 621–638, 2014.

Harold W. Kuhn. A simplified two-person poker. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games, Volume I*, volume 24 of *Annals of Mathematics Studies*, pp. 97–103. Princeton University Press, 1950.

Marc Lanctot, Kevin Waugh, Martin Zinkevich, and Michael Bowling. Monte carlo sampling for regret minimization in extensive games. *Advances in neural information processing systems*, 22, 2009.

Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. Openspiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019.

Boning Li and Longbo Huang. Efficient online pruning and abstraction for imperfect information extensive-form games. In *The Thirteenth International Conference on Learning Representations*, 2025.

Boning Li and Longbo Huang. Effective, efficient, and general information abstraction for imperfect-information extensive-form games. *arXiv preprint arXiv:2605.10900*, 2026.

Boning Li, Zhixuan Fang, and Longbo Huang. Rl-cfr: improving action abstraction for imperfect information extensive-form games with reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 27752–27770, 2024.

Hsieh-Ting Lin and Tsung-Yu Hou. Readable minds: Emergent theory-of-mind-like behavior in llm poker agents. *arXiv preprint arXiv:2604.04157*, 2026.

Zachary C Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43, 2018.

Edward Lockhart, Marc Lanctot, Julien Pérolat, Jean-Baptiste Lespiau, Dustin Morrill, Finbarr Timbers, and Karl Tuyls. Computing approximate equilibria in sequential adversarial games by exploitability descent. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 464–470, 2019.

Runyu Lu, Yuanheng Zhu, and Dongbin Zhao. Constrained exploitability descent: An offline reinforcement learning method for finding mixed-strategy nash equilibrium. In *Forty-second International Conference on Machine Learning*, 2025.

Carlos Martin and Tuomas Sandholm. Approxed: Approximate exploitability descent via learned best responses. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pp. 1454–1463, 2025.

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 627–635, 2011.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Andrei A Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. In *International Conference on Learning Representations*, 2016.

Tuomas Sandholm. Abstraction for solving large incomplete-information games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Martin Schmid, Matej Moravčík, Neil Burch, Rudolf Kadlec, Josh Davidson, Kevin Waugh, Nolan Bard, Finbarr Timbers, Marc Lanctot, G Zacharias Holland, Elnaz Davoodi, Alden Christianson, and Michael Bowling. Student of games: A unified learning algorithm for both perfect and imperfect information games. *Science Advances*, 9(46):eadg3256, 2023.

Finnegan Southey, Michael Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and Chris Rayner. Bayes' bluff: opponent modelling in poker. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pp. 550–558, 2005.

Umar Syed, Michael Bowling, and Robert E Schapire. Apprenticeship learning using linear programming. In *Proceedings of the 25th international conference on Machine learning*, pp. 1032–1039, 2008.

Oskari Tammelin, Neil Burch, Michael Johanson, and Michael Bowling. Solving heads-up limit texas hold'em. In *IJCAI*, volume 15, pp. 645–652, 2015.

Luís Filipe Teófilo and Luís Paulo Reis. Building a no limit texas hold'em poker agent based on game logs using supervised learning. In *International Conference On Autonomous And Intelligent Systems*, pp. 73–82. Springer, 2011.

Luís Filipe Teófilo, Luís Paulo Reis, Henrique Lopes Cardoso, and Pedro Mendes. Rule based strategies for large extensive-form games: A specification language for no-limit texas hold'em agents. *Computer Science & Information Systems*, 11(4), 2014.

Finbarr Timbers, Nolan Bard, Edward Lockhart, Marc Lanctot, Martin Schmid, Neil Burch, Julian Schrittwieser, Thomas Hubert, and Michael Bowling. Approximate exploitability: Learning a best response. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 3487–3493. ijcai.org, 2022. doi: 10.24963/ijcai.2022/484.

Kevin Waugh, David Schnizlein, Michael H Bowling, and Duane Szafron. Abstraction pathologies in extensive games. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems*, volume 2, pp. 781–788, 2009.

Guan Yang, Minghuan Liu, Weijun Hong, Weinan Zhang, Fei Fang, Guangjun Zeng, and Yue Lin. Perfectdou: Dominating doudizhu with perfect information distillation. *Advances in neural information processing systems*, 35:34954–34965, 2022.

Brian Hu Zhang and Tuomas Sandholm. Small nash equilibrium certificates in very large games. In *NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4fbe073f17f161810fdf3dab1307b30f-Abstract.html.

Brian Hu Zhang and Tuomas Sandholm. Finding and certifying (near-) optimal strategies in black-box extensive-form games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 5779–5788, 2021a.

Brian Hu Zhang and Tuomas Sandholm. Subgame solving without common knowledge. *Advances in Neural Information Processing Systems*, 34:23993–24004, 2021b.

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. *Advances in neural information processing systems*, 20, 2007.

A Extended related work

Exact evaluation. Our evaluation is an exact best-response computation over the full game, so each reported NashConv value evaluates that instance exactly.

B Derivation of Example 1

Player 1's shared leaf plays a_1 with probability q in both states. Against the best-responding opponent, who punishes whenever punishing is profitable, player 1's expected payoff is

$$u_1(q) \;=\; (1-\epsilon)\,q\,\delta \;+\; \epsilon\,q\,(-M) \;=\; q\big[(1-\epsilon)\delta - \epsilon M\big]. \tag{6}$$

The unconstrained optimum for player 1 in the full game is $(1-\epsilon)\delta$, achieved by playing a_1 in A and a_2 in B, so the exploitability gap of the leaf is

$$g(q) \;=\; (1-\epsilon)\delta - u_1(q) \;=\; (1-\epsilon)\delta\,(1-q) \;+\; q\,\epsilon M. \tag{7}$$

The imitation objective is the reach-weighted forward KL to the teacher, minimized by the reach-weighted average of the teacher's actions, $q_{\mathrm{im}} = 1-\epsilon$, giving

$$g(q_{\mathrm{im}}) \;=\; \epsilon(1-\epsilon)\,(\delta + M). \tag{8}$$

The exploitability objective minimizes g, which is linear in q: when $\epsilon M > (1-\epsilon)\delta$ the minimizer is $q_{\mathrm{ex}} = 0$ with $g(q_{\mathrm{ex}}) = (1-\epsilon)\delta$. The ratio is

$$\frac{g(q_{\mathrm{im}})}{g(q_{\mathrm{ex}})} \;=\; \frac{\epsilon\,(\delta + M)}{\delta}, \tag{9}$$

unbounded as $M \to \infty$ at any fixed $0 < \epsilon < 1$ and $\delta > 0$. For instance $\epsilon = 0.05$, $\delta = 0.1$, $M = 10$ gives ratio 5.05, and $M = 1000$ gives ratio 500.05. The derivation uses the unfloored forward KL; the 10^{-12} floor of Section 3 moves the reach-weighted average by at most 10^{-12} and changes none of the displayed figures.

C Proof of Proposition 1

Fix p and a pure best response μ_p of player $1-p$ to σ_p; a pure best response exists for any fixed opponent profile in a finite perfect-recall game. Write $u(\cdot)$ for player $(1-p)$'s expected utility. Because μ_p is a best response to σ_p and some strategy is a best response to τ_p,

$$\mathrm{BR}_{1-p}(\sigma_p) - \mathrm{BR}_{1-p}(\tau_p) \;=\; u(\sigma_p, \mu_p) - \max_{\mu'} u(\tau_p, \mu') \;\le\; u(\sigma_p, \mu_p) - u(\tau_p, \mu_p). \tag{10}$$

The right side compares two profiles that differ only in player p's strategy. Order p's information sets so that no set precedes one of its ancestors and switch p from τ_p to σ_p one set at a time. When set I is switched, everything above I already follows σ_p, so I is reached with p's own probability $\pi_p^{\sigma}(I)$, and everything below I still follows τ_p; the histories in I carry the chance-and-opponent reach $\pi_{-p}^{\mu_p}(h)$, which is independent of p's strategy. Let $q_I(a)$ be player $(1-p)$'s counterfactual value of action a at I in this intermediate profile, normalized by $\pi_{-p}^{\mu_p}(I) = \sum_{h\in I} \pi_{-p}^{\mu_p}(h)$; by perfect recall the values below I depend on p's strategy only through the sets that follow I, all still at τ_p, so q_I is the quantity defined in the proposition. The change in u from switching I is $\pi_p^{\sigma}(I)\,\pi_{-p}^{\mu_p}(I)\sum_a(\sigma_p(I,a) - \tau_p(I,a))\,q_I(a)$, and the switches telescope to $u(\sigma_p,\mu_p) - u(\tau_p,\mu_p)$. Finally, for any two distributions x, y and any vector q with range r, $\sum_a (x_a - y_a) q_a \le r\,\mathrm{TV}(x,y)$, since the sum is unchanged when a constant is subtracted from q and the positive part of $x - y$ has mass $\mathrm{TV}(x,y)$. Summing over p and applying Lemma 1 gives (5); $r_I \le \Delta$ because every $q_I(a)$ is an average of terminal utilities.

The identity behind the telescoping step and the inequality are checked numerically on Leduc: on the aligned $B = 8$ structure with best-found leaves, the per-player telescoped sum matches $u(\sigma_p,\mu_p) - u(\tau_p,\mu_p)$ to 10^{-9}, and the NashConv gap 0.867 sits below the bound 4.29 with the local ranges r_I (53.97 with r_I replaced by $\Delta = 26$). With imitation leaves on the same structure the gap 4.440 sits below 8.13 (56.73). On 60 random students spanning pure, Dirichlet, and teacher-mixture

profiles the inequality never fails and the local bound comes within a factor 1.03 of the true gap in the tightest case. Normalizing the reach products $\pi_p^{\sigma}(I)\,\pi_{-p}^{\mu_p}(I)$ (without r_I) and the imitation weights $\pi_{-p}^{\tau}(I)$ to distributions over $\mathcal{I}_p$, the two differ by total variation 0.34 and 0.23 for players 0 and 1 (0.33 and 0.28 for the imitation-leaf student), and 17 and 13 percent of the bound's reach weight (37 and 17 percent) falls on information sets that hold under one percent of the imitation weight.

D READABLE STRATEGIES, WORKED EXAMPLES

We make the readability claim of Section 5 concrete on the aligned $B = 8$ tree of Table 3. Every rule tests one routing condition (whether a bet is faced, or the raise cap is reached) and at most two card conditions, and one action distribution is shared by all information sets of a leaf. The rules are first-match in the order listed: the first rule absorbs every information set where no bet is faced *or* the raise cap has been reached, so the three "facing a bet" rules that follow apply only to a bet below the cap. Rebuilding the student from the eight rules reproduces the reported NashConv to 10^{-12}, and rounding each leaf to its most likely action (the two raise leaves move from 0.993 and 0.973 to 1) leaves both best-response values unchanged at exactly $13/30$ for the rounded pure student (the mixed student evaluates to 0.8666666666680334), so the pure rule list has the same exploitability as the mixed student. Three situations, covering three of the four player-0 leaves, show what a person reads and what the student plays next to the teacher.

1. *Round one, first action, player 0 holds a low card.* The leaf covers all 96 two-action information sets, among them the opening position with a low card. The student checks with probability 1.00; the teacher checks with probability 0.928 and raises with probability 0.072 on the same information set. The student drops the occasional raise of a weak hand that the teacher keeps.
2. *Round one, player 0 holds the two and faces a bet.* The rule $card = 2$ and $board \leq -1$ fires, a leaf that covers exactly this one information set. The student raises with probability 0.993; the teacher calls with probability 0.318 and raises with probability 0.682. Two conditions decide the whole play, and the mix is compressed to a near-constant bet.
3. *After the flop, player 0 holds the two and faces a bet.* The rule $card = 2$ and $board > -1$ fires and covers 15 information sets. The student calls with probability 1.00 on all of them; of the two representative information sets the teacher folds with probability 1.00 on one and calls with probability 1.00 on the other. The student replaces 15 separate teacher distributions with one always-call rule.

A person simulating the student reads one routing condition, at most two card conditions, and one of eight rules, and returns the same action distribution for a whole group of information sets, so the play is reproducible from a short sheet. This is the read-and-simulate property the Introduction requires of a distilled strategy: the student is a small object whose decisions a person can walk and execute without consulting the teacher's per-information-set table.

E IMPLEMENTATION CHECKS

Two identities are asserted programmatically in every run. First, Lemma 1: the exact NashConv of the lifted student profile equals the sum of the two per-player objectives $\mathrm{BR}_0(\sigma_1) + \mathrm{BR}_1(\sigma_0)$, checked to exact floating-point equality at every budget of the factorial scan. Second, the scalarized identity $f_0^{\eta} + f_1^{\eta} = \mathrm{NashConv} + \eta(F_0 + F_1)$ of Section 4, which guards against the objective silently dropping the fidelity term; at $\eta = 0$ the scalarized solver reproduces the pure exploitability solver to exact floating-point equality on all four budgets, per seed. Two degenerate cases have fixed conventions: a leaf whose information sets all carry zero teacher-opponent reach is fit to the uniform distribution, and F_p is zero when the reach sum over player p's information sets is zero.

The structure-overlap metrics of Section 5 are defined as follows. Each information set maps to the ordered tuple of split conditions on its root-to-leaf path. Path Jaccard is the Jaccard index of the two sets of distinct condition tuples. Split-feature Jaccard at growth step d is the Jaccard index of the sets of feature names used by the first d splits of each arm. The partition metrics compare leaf assignments. Each information set of a player carries its leaf label; co-membership Jaccard is the

Jaccard index of the two sets of information-set pairs that share a leaf, and the Rand index is the fraction of pairs on which the two assignments agree, whether they share a leaf or not. All four metrics are computed by re-growing the lifted and KL trees under the run protocol and comparing them per seed; at $B = 8$ the co-membership Jaccard is 0.456 to 0.679 and at $B = 16$ it is 0.373 to 0.561, with Rand indices of 0.85 to 0.92 at both budgets.

The same path-Jaccard index applied to the aligned arm gives the overlap quoted in Section 4: against the lifted-grown trees its maximum over the 8 seeds is 0.44 at $B = 8$ and 0.07 at $B = 16$, and against the KL-grown trees 0.27 and 0.07; at $B = 4$ the three arms share one tree and the index is 1 for every seed.

The best-response evaluator behind every NashConv in this paper is cross-checked against an independent implementation. On Leduc the 288 local information sets map one-to-one onto those of OpenSpiel (Lanctot et al., 2019), with no unmatched information set, and the two evaluators agree to a relative error of 10^{-13} or better on the uniform profile, on a 3000-iteration CFR+ profile, and on three random profiles. Kuhn's uniform profile evaluates to $11/12$ in both.

F KUHN PROTOCOL AND TEACHER

Kuhn poker has 12 information sets in our encoding, three cards, one betting round, and a one-chip ante. The handcrafted grouping used in Section 5 maps each information set to its card and the parity of the betting history, which yields six leaves, three per player. The three quoted numbers come from one teacher: plain CFR for 20000 iterations, average strategy returned, NashConv 7.5×10^{-3}. On the six handcrafted leaves, grouped CFR+ for 800 iterations, warm-started from the imitation fit, returns NashConv 0.364, and the number is stable across iteration counts: at 200, 800, 3200, and 6400 iterations the same solve returns 0.3644, 0.3635, 0.3639, and 0.3639, a spread of 9×10^{-4}; the best-found leaf solve of Equation (3) on the same six leaves returns 0.138; and best-first growth guided by the exploitability score to six leaves with the same leaf solve returns 0.056. The same three protocols under the CFR+ teacher of Section 5 (3000 iterations, NashConv 7.3×10^{-5}) return 0.364, 0.146, and 0.091, so the handcrafted-versus-learned gap persists at a much stronger teacher.

G ROBUSTNESS OF THE ALIGNED SCORE

Teacher strength: rerunning the full aligned pipeline at $B \leq 8$ with teachers trained for 1000, 3000, and 6000 CFR+ iterations leaves the final best-found NashConv identical in every bit (0.8666666666680334 at $B = 8$), although the grown trees differ slightly in their imitation-fitted diagnostics.

Teacher-free starts and arm. The five-start leaf solve contains three starts that never consult the teacher (uniform and two fixed-seed Dirichlet draws). Across the best-found arms of the main Leduc and five-rank experiments, a teacher-free start supplies the winning per-player solve in 41 of 224 cases (18.3 percent), which bounds how often the teacher-seeded start wins inside the final leaf solve.

The end-to-end arm goes further: the teacher is replaced by the uniform profile (NashConv 4.747 on Leduc, 4.858 on the five-rank variant) in all of its roles (the warm leaf, the imitation start, the reach weights, and the reference the split score reads), and no teacher solve is run. The leaf solve keeps five starts under the same solver settings, the imitation start replaced by a third Dirichlet draw (CFR, uniform, three Dirichlet draws), so the per-candidate cost matches the teacher-guided arm, and every other setting is the one the mirror run used; the only saving is the teacher solve itself. On Leduc the substitution changes nothing at $B \leq 8$: all 8 seeds return the same NashConv as the teacher-guided arm to machine precision (2.000000 and 0.866667), that is, the same deterministic structure. At $B = 16$ the two arms differ by seed, as any tolerance-randomized protocol does, and the teacher-free arm is lower on 5 of 8 seeds with IQM 0.792 against 0.807. On the five-rank variant at $B = 8$, under the all-candidate protocol, the teacher-free arm returns the teacher-guided value on 4 of 8 seeds and 0.028 to 0.032 above it on the other 4 (IQM 0.693 against 0.678). The teacher therefore supplies a warm start. The aligned split score is a full-game best-response objective, and on Leduc it selects the same structure with or without it (Table 4).

Table 4: Teacher-free arm against its teacher-guided mirror, NashConv. The mirror is the run whose settings the teacher-free arm copies exactly; both arms run five leaf-solve starts under the same solver settings, the teacher-free arm replacing the imitation start by a third Dirichlet draw. Per-seed values are identical to machine precision on the first two rows, so those rows quote one value.

game	arm	teacher-guided	teacher-free	lower/equal/higher
Leduc, $B = 4$	all candidates	2.000000	2.000000	0/8/0
Leduc, $B = 8$	all candidates	0.866667	0.866667	0/8/0
Leduc, $B = 16$	all candidates	0.807 (IQM)	0.792 (IQM)	5/0/3
Leduc(5,2), $B = 8$	all candidates	0.677778	0.693 (IQM)	0/4/4

Table 5: Leduc at $B = 16$: per-seed NashConv of the aligned arm and the KL arm under the full best-found leaf solve. Every aligned seed lies below every KL seed. Aligned seeds 2 to 7 are cluster runs (Appendix L).

	seed							
arm	0	1	2	3	4	5	6	7
aligned	0.805	0.817	0.841	0.807	0.814	0.787	0.803	0.763
KL split	0.892	1.007	0.858	0.939	1.000	0.892	0.967	0.934

Score tolerance: tolerances of 1, 5, and 10 percent produce the identical tree at $B \leq 8$ for the aligned score, because the winning candidate leads the runner-up by roughly 35 percent, far above the tolerance window; the per-seed draw is taken at every split but has a single candidate to pick from, and all 8 seeds grow the same tree. Both facts sharpen the statistical reading of Table 2: at $B \leq 8$ the aligned arm's spread across seeds reflects only leaf-solver floating point noise, so we report its result as one deterministic structure, and the 8/8 win count compares that structure against 8 genuinely distinct trees per baseline arm. At $B = 16$ the tolerance window does begin to bind and seeds diverge. All 8 seeds (0.763 to 0.841) lie below all 8 KL-score seeds, whose minimum is 0.858, and the 8 values are distinct, so this is an 8-versus-8 separation rather than a single deterministic structure (Table 5). The eight seeds induce eight distinct NashConv values at $B = 16$, and the six growth states the run retained induce six distinct partitions, one per seed; at $B \leq 8$ the retained states all induce the same partition, which is why the smaller budgets report one structure.

The prefilter sweep of Table 6 localizes the loss. With the aligned fine score applied only to the top K candidates of the lifted prefilter, $K \leq 8$ puts every seed on the value of the zero-tolerance lifted-score tree, and the gain returns only once the shortlist grows past the point where the mismatched ranking decides the outcome.

H Cost accounting

All timings are single-core CPU measurements from the run logs unless noted. Candidate evaluation: about 10 to 20 ms for imitation-scored candidates and about 20 s for aligned-scored candidates (one Powell start from the imitation fit, capped iterations). The first Leduc split ranks 64 candidates; later steps grow with the leaf count. End-to-end growth per seed with the aligned score: about 74 minutes to $B = 8$; between 4.5 and 10 hours to $B = 16$ depending on solver trajectories. The imitation prefilter with $K = 8$ reduces each step to 8 aligned evaluations plus the closed-form ranking pass. On the five-rank variant the prefiltered aligned arm needs 68 to 99 minutes of single-core time to reach $B = 8$ (60 to 89 minutes of growth plus the teacher solve) and 6.5 to 9 hours to $B = 16$. Scoring every candidate there takes 2,166 to 2,337 seconds of growth per seed on a 15-worker pool at $B = 8$, that is 36 to 39 minutes of wall time and about 9 to 10 CPU hours, plus a 178 to 192 second teacher solve. The end-to-end wall time per seed is 39 to 42 minutes, excluding the final five-start leaf solve described below. That pool occupies 15 cores while the prefiltered numbers above are single-core, so the two are not on the same axis; a dedicated one-core rerun of seed 0 puts the all-candidate growth to $B = 8$ at 16,485 seconds (4.6 hours) of single-core time, three to five times the prefiltered arm's 60 to

Table 6: Leduc at $B = 8$, full best-found leaf solve: the aligned fine score applied to the top K candidates of the lifted prefilter. At $K \leq 8$ every seed lands on the zero-tolerance lifted-score value; at $K = 16$ one seed recovers part of the gain.

	seed							
K	0	1	2	3	4	5	6	7
2	1.533	1.533	1.533	1.533	1.533	1.533	1.533	1.533
4	1.533	1.533	1.533	1.533	1.533	1.533	1.533	1.533
8	1.533	1.533	1.533	1.533	1.533	1.533	1.533	1.533
16	1.533	1.520	1.533	1.533	1.478	1.067	1.533	1.528
zero-tolerance lifted score	1.533							

Table 7: The five-rank variant at $B = 8$, full best-found leaf solve, three split protocols. The all-candidate aligned arm grows the same partition from all 8 seeds (per-seed values within 5×10^{-14}), so its row is a single deterministic value; the other two rows are 8 distinct trees. The all-candidate arm is a cluster run (Appendix L).

	seed							
arm	0	1	2	3	4	5	6	7
KL split	0.771	1.028	0.729	0.811	0.704	0.798	0.771	0.791
aligned, $K = 8$ prefilter	0.811	0.861	0.811	0.811	1.406	0.811	0.811	0.861
aligned, all candidates	0.678 (same partition in all 8 seeds)							

89 minutes, and reproduces its NashConv 0.678; Figure 4 itself shows only the prefiltered arm. The full five-start leaf solve used for reported numbers costs between tens of seconds and a few minutes per budget point at $B \leq 16$.

I THE FIVE-RANK VARIANT

Leduc(5,2) has 780 information sets, 2.7 times Leduc, and runs the same protocol with the same 8 seeds. Figure 4 shows the $K = 8$ prefiltered aligned arm, because fine-scoring every candidate at every step costs far more here than on Leduc. Table 7 adds the all-candidate arm at $B = 8$.

The all-candidate seeds are not independent draws: all 8 end at the same 8-leaf partition, and their NashConv values differ by at most 5×10^{-14}, so the table reports one deterministic value. They do not share the whole path. Candidate scores are seed-independent (one imitation-initialized solve per candidate under a fixed solver seed), so a seed can only matter when the tolerance pool holds more than one candidate, and it does at two of the four steps. At $B = 4 \to 5$ seeds 1 and 6 take a different split from the other six; at $B = 6 \to 7$ seeds 1, 5, and 6 do; and all eight rejoin one partition at $B = 6$ and again at $B = 8$. These are the six partitions of Section 5. The prefiltered arm, by contrast, produces 8 distinct trees and 8 distinct values, including one seed at 1.406.

J QUALITY-AWARE LEAF FITTING

The imitation leaf fit minimizes a reach-weighted KL, and the reach weights are not the teacher's advantage: an error at an information set that a best responder punishes hard carries no extra penalty. We test whether that weighting accounts for the gap of Table 1 by fitting the same closed-form teacher average under VIPER-style risk weights (Bastani et al., 2018). For each player we normalize the teacher's counterfactual action values, $q_n(I, a) = q(I, a)/\pi^{\sigma_T}_{-p}(I)$, and weight information set I by $w(I) = \max(0, V(I) - \min_a q_n(I, a))$ with $V(I) = \sigma_T(I)^\top q_n(I)$; the weighted fit is the same reach-weighted average with $w(I)$ folded into the reach. A second arm weights by the per-information-set regret $\max_a q_n(I, a) - V(I)$ instead.

Table 8: Quality-aware weighted leaf fits on the same trees as Table 1 (Leduc, IQM over 8 seeds, exact NashConv). VIPER and regret weight the same closed-form teacher average by the teacher's own advantage and by its per-information-set regret; best-found is the leaf solve of Table 1. Every weighted cell is more exploitable than the unweighted fit it replaces.

Score	B	imitation	VIPER	regret	best-found	ratio	VIPER ratio
lifted	4	4.644	5.004	4.696	2.000	2.32	2.50
lifted	8	3.954	4.336	4.943	1.533	2.58	2.83
lifted	16	3.096	3.527	4.597	1.065	2.91	3.31
lifted	32	2.321	2.956	4.791	0.913	2.54	3.24
kl	4	4.644	5.004	4.696	2.000	2.32	2.50
kl	8	4.461	4.862	5.095	1.352	3.30	3.60
kl	16	4.026	4.535	4.429	0.933	4.32	4.86
kl	32	3.011	3.468	3.987	0.777	3.88	4.47

Table 8 reports both arms on identical trees, budgets, and seeds, with the imitation and best-found columns reproduced from Table 1; the imitation rows were re-run and match the factorial scan bit for bit. Reweighting does not close the gap. The VIPER-weighted fit is worse than the unweighted fit it replaces in all eight cells, raising the ratio from 2.32 to 2.50 at $B = 4$ and from 2.91 to 3.31 at $B = 16$ under the exploitability split score, and its reach-weighted fidelity gap is wider (1.125 against 1.031 at $B = 4$). The regret-weighted arm is weaker still, and its weights are degenerate on a teacher this close to equilibrium: over the 288 information sets the median weight is 1.0×10^{-5}, three quarters fall below 10^{-4} (maximum 3.92), and one leaf in one cell ends with zero total weight and falls back to uniform. The objective, not the weighting, separates the arms.

K A SECOND GAME: LIAR'S DICE

Liar's Dice with one four-sided die per player tests whether the ordering of Section 5 survives outside the Leduc family. Two players alternate bids of the form *quantity, face* and may challenge instead of raising; a challenge is resolved on the two hidden dice. Our encoding follows OpenSpiel's topology and its wild rule (Lanctot et al., 2019), under which the highest face counts towards every bid: with the uniform profile in both roles the two encodings return the same NashConv to the last bit, 1.3101190476190476. The game has 1024 information sets and 4080 terminal histories, and its action-count buckets give 16 initial leaves, so the three budgets are the initial tree (16), 24, and 32. The protocol is the one of Section 5: the same growth rules and one-percent tolerance, the same eight seeds, and a CFR+ teacher at NashConv 3.9×10^{-5}, identical in all eight runs.

Table 9 repeats the factorial comparison on these trees. The best-found leaf solve is below the imitation fit on every seed in all six structure-budget cells, by factors of 1.6 to 3.3 on the interquartile means. The grouped CFR+ baseline of Section 5, run on the exploitability-scored structures, is worse than the best-found solve in every cell and beats the imitation fit only at $B = 16$ (0.875 against 0.887), where no growth step runs and all eight seeds share the initial tree. At $B = 24$ and $B = 32$ it is worse than the imitation fit on its own structure (0.652 against 0.445 and 0.483 against 0.208), so solving the compressed game does not recover what the leaf fit loses. The last column of the table records how many distinct values the eight seeds produce in each cell.

L REPRODUCIBILITY

An anonymized copy of the code is included with this submission: the game engine, the teacher, the exact evaluation, and the drivers for the main experiments and their baselines, compiled into binary modules for Linux x86-64, with a short launcher and a requirements file. This section records the implementation details needed to reproduce the main experiments.

Games. Leduc poker (Southey et al., 2005) uses a six-card deck (three ranks, two suits), one private card per player, a one-chip ante, and one public board card dealt before the second round. There are two betting rounds with at most two raises per round, player 0 acts first in both, and the legal

Table 9: Factorial comparison on Liar's Dice 1×4, with the arms and budgets of Table 1 (IQM over 8 seeds, exact NashConv on the original game). The ratio is imitation over best-found; the last column counts how many distinct values the eight seeds produce in the imitation and the best-found arm, so $1/1$ records one distinct value; it holds exactly at $B = 16$, where no growth step runs and all eight seeds share the initial tree by construction.

Score	B	imitation	best-found	ratio	distinct
lifted	16	0.8870	0.5388	1.65	$1/1$
lifted	24	0.4454	0.1458	3.05	$6/8$
lifted	32	0.2075	0.0625	3.32	$8/8$
kl	16	0.8870	0.5388	1.65	$1/1$
kl	24	0.6624	0.4104	1.61	$2/8$
kl	32	0.5152	0.1875	2.75	$3/8$

Table 10: Implementation parameters; rows whose setting differs for a specific experiment say so.

Component	Setting
Leaf solver	Powell's method, function tolerance 10^{-9}, iteration cap 800, evaluation cap 6400
Multi-start set	imitation fit, restricted-game CFR+ (1000 iterations, warm-started from the imitation fit), uniform, and two fixed-seed Dirichlet draws; the best of all starts is returned
Aligned split scoring	one Powell start from the imitation fit, iteration cap 200, no CFR, no Dirichlet
Probability parameterization	free logits with the last entry pinned to zero; logits shifted by the maximum, clipped to $[-30, 0]$, then softmax. The imitation start is floored at 10^{-12} before taking logits
Grow tie-breaking	with a one-percent tolerance, candidates within one percent of the best score tie-break uniformly at random on a per-seed stream; with zero tolerance the first minimum in enumeration order wins
Teacher	CFR+ for 3000 iterations, deterministic and recomputed by every run, so the two machines we used return different teachers; the Kuhn concept check uses plain CFR for 20000 iterations (Appendix F)
Restricted-game baseline	grouped CFR+ for 800 iterations on the lifted structure, warm-started from the reach-weighted average; a restricted-game heuristic, defined below
Split prefilter	top $K = 8$ candidates by the lifted score; used for the five-rank prefiltered arm of Figure 4 and for the counterfactual sweep of Table 6. Every other arm, including the five-rank all-candidate arm, scores every candidate
Fidelity weight	η swept over 11 values, 0 to 1000, in the frontier experiment
KL clipping	10^{-12} floor before both renormalization steps
Split candidates	ordinal threshold splits ($\leq$) and categorical equality splits, each checked for a uniform action count on both sides

actions are check, call, raise, and fold. A raise first matches the outstanding difference and then adds the round's bet size, two chips in round one and four in round two; a fold concedes the folder's own contribution. At showdown a player whose private card pairs the board wins, otherwise the higher private rank wins and equal ranks tie, and the winner receives the loser's contribution, the two being equal at showdown. With R ranks the game has $6R + 30R^2$ information sets in our encoding, 288 at $R = 3$ (Leduc) and 780 at $R = 5$ (the five-rank variant), keyed by the acting player's private card, the public board, and the betting history. Kuhn poker (Kuhn, 1950) uses three cards, one private card each, a one-chip ante, and one betting round with a one-chip bet; its 12 information sets are keyed by card and history.

Hyperparameters. Table 10 lists the parameters of the main experiments; rows whose setting differs for a specific experiment say so.

Growth protocols. The three split scores share one loop and differ only in how a candidate split is scored. Each run starts from the coarsest legal tree, one leaf per player and action count. At every step the loop enumerates the candidate splits of the current tree, scores each candidate, applies the best-scoring split, and records the intermediate tree. It stops at the budget B, so the intermediate trees at every leaf count form one incremental path.

1. *Lifted score.* For each candidate, fit the leaves by the reach-weighted teacher average and compute the exact NashConv of the tentative tree, Equation (1). The candidate is scored by that NashConv. This is the exploitability-aware baseline labeled expl. in the figures.
2. *KL score.* For each candidate, fit the leaves as above and compute the symmetric KL from the teacher to the tentative tree, defined below. The candidate is scored by that distance.
3. *Aligned score.* For each candidate, fit the leaves with a cheap best-found solve of the leaf objective Equation (3), one Powell start from the imitation fit and an iteration cap of 200, and score the candidate by the NashConv the solve reaches. The optional two-stage variant first ranks all candidates by the lifted score and passes only the top $K = 8$ to the aligned fine score.

In all three protocols the winning candidate follows the one-percent tolerance rule of Table 10: the best score s^* is the minimum over candidates, and the applied split is drawn uniformly at random among candidates with score at most $(1 + 0.01)\, s^*$, using a per-seed random stream. The zero-tolerance deterministic baselines of Table 2 run the same loop with the tolerance set to zero, which selects the first minimum in enumeration order.

Restricted-game baseline. Grouped CFR+ treats each shared leaf as one decision variable. At every alternating player pass it lifts the current leaf strategy to a full-game profile, computes the counterfactual action values $q_I(a)$ at each information set, forms the counterfactual gap $q_I(a) - \sum_{a'} \sigma_b(a') q_I(a')$ for the leaf b that owns I, sums those gaps over all information sets routed to b, adds the sum to b's clipped regret vector, and regenerates the leaf distribution by regret matching, which plays each action in proportion to its positive clipped regret. The average strategy is reach-weighted with linear averaging, as in CFR+. Leaf sharing can merge information sets that differ in the acting player's own past actions, so perfect recall need not survive the grouping and the CFR+ guarantee does not transfer to the grouped update. We therefore treat this arm as a restricted-game heuristic and read its numbers as one heuristic way to solve the compressed game. The comparison it supports is between the compressed-game objective and the full-game exploitability we evaluate, and the gap confounds that objective difference with the heuristic's own optimization quality, so we report the arm as a baseline and attribute the gap to neither alone.

KL clipping. The split score uses the symmetric KL between two action distributions a and b,

$$\mathrm{KL}(a, b) \;=\; \sum_i a_i \log \frac{a_i}{b_i}, \tag{11}$$

evaluated in both directions and averaged, without reach weights, over the teacher's information sets. Both inputs are clipped to 10^{-12} and renormalized before the sum, so a zero probability never produces a logarithm of zero. The floor still matters when the teacher is exactly deterministic: a student that puts positive mass where the teacher does not is charged $\log(10^{12})$ in the reverse direction, so the floor bounds that term without removing it. The fidelity term of Equation (4) uses the forward reach-weighted KL,

$$F \;=\; \frac{\sum_I \mathrm{reach}(I)\, \mathrm{KL}\big(\sigma^\star(I) \,\|\, \sigma(I)\big)}{\sum_I \mathrm{reach}(I)}, \tag{12}$$

with the same 10^{-12} clipping and renormalization, where information sets with reach at most zero are dropped from both sums.

The unweighted form is the better split score, and the simpler one. The reach-weighted forward KL that the imitation fit itself minimizes is the natural alternative, since it is the loss the leaf fit optimizes; as a split score it is no better at $B = 4$ and worse from $B = 8$ under both leaf solves (Table 11). That comparison changes two things at once, the reach weighting and the KL direction, so we read it as a statement about two complete scoring rules: matching the split loss to the leaf fit does not rescue the imitation-oriented ranking.

Table 11: Reach-weighted forward KL, the score the imitation leaf fit minimizes, against the unweighted symmetric KL as split scores on Leduc, IQM over 8 seeds under both leaf solves. The unweighted score ties at $B = 4$ and is lower at every larger budget. The reach-weighted rows are cluster runs (Appendix L).

leaf solve	split score	$B = 4$	$B = 8$	$B = 16$	$B = 32$
imitation	reach-weighted forward KL	4.644	4.814	4.507	3.972
	unweighted symmetric KL	4.644	4.461	4.026	3.011
best-found	reach-weighted forward KL	2.000	1.468	1.298	0.950
	unweighted symmetric KL	2.000	1.352	0.933	0.777

Reach weights and teacher. The weight $\mathrm{reach}(I)$ of an information set is the counterfactual reach of the opposing player: it multiplies the chance probabilities and the opposing player's action probabilities, but never the action probabilities of the player whose information set is being weighted. A leaf's imitation strategy is the reach-weighted average of the teacher within that leaf, renormalized by the leaf's total reach; leaves with zero total reach fall back to the uniform distribution. The teacher is a deterministic CFR+ run of 3000 iterations, recomputed by every run. The same specification returns different teachers on the two machines we used: NashConv 1.37×10^{-4} on Leduc and 2.151×10^{-3} on the five-rank game on the workstation, 6.79×10^{-4} and 2.153×10^{-3} on the cluster nodes. Cluster runs are: the aligned arm's seeds 2 to 7 at $B = 16$ (Table 5; the $B = 16$ aligned row of Table 2 mixes the two teachers; the six structures of the aligned frontier are these six seeds), the reach-weighted rows of Table 11, and the all-candidate arm of Table 7. All other rows use the workstation teachers. The Leduc comparisons built on the cluster runs pair arms grown under teachers that differ by a factor of five; on the five-rank game the two teachers agree to 0.1 percent. Regrowing seeds 2 and 7 end to end on the workstation reproduces NashConv 2.000000 and 0.866667 at $B = 4$ and 8 exactly and returns 0.8333 and 0.8099 at $B = 16$ against the stored 0.8408 and 0.7632: a different path at that budget, with both values still below the KL arm's minimum 0.858. Within one machine, seed-to-seed variation reflects only the tolerance-based random tie-breaking of the structure search.

Compute-matched restart control. The restart control of Section 5 gives the KL score, per seed, the CPU budget the aligned arm spent on the same seed, measured as process CPU time so that machine load cannot distort the comparison. The budget is divided into R independent KL growths at $B = 8$, each with its own tie-breaking stream (stream $1000s + r$ for seed s and restart r) under the same one-percent tolerance; R is calibrated per seed from single-restart CPU cost and lands at 153 to 269 on Leduc and 10 to 16 on the five-rank variant. Each grown tree receives the cheap aligned-scoring solve of Table 10 as a proxy, the tree with the best proxy NashConv wins, and only the winner receives the full multi-start solve used by every other arm. Every seed finished within its budget (at most 88 percent used on Leduc and 78 percent on the five-rank variant). On Leduc all eight seeds return the same final value, 0.9667, but the winning trees were not retained, so we report eight independent growths that agree in value and make no claim that they share a structure.

Frontier sweep. The scalarized frontier of Figure 3 sweeps $\eta \in \{0, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 100, 1000\}$ at $B \in \{4, 8, 16, 32\}$ on one fixed structure, the exploitability-grown tree of the factorial scan, with the full best-found leaf solve, 352 cells in all. The fidelity sum $F_0 + F_1$ at $\eta = 0$ is 28.1, 25.7, 22.5, and 12.5 (IQM over seeds) at the four budgets, and 6.5, 5.6, 4.2, and 2.7 at $\eta = 0.01$, so the first grid step buys a 4.3- to 5.4-fold drop in fidelity gap for an IQM NashConv rise of at most 0.016. Across the 320 consecutive grid pairs (8 seeds $\times$ 4 budgets $\times$ 10 pairs) NashConv never decreases and $F_0 + F_1$ never increases, with zero violations of either direction.

The floor gives those numbers their scale. On a fixed grouping the reach-weighted forward KL is minimised in closed form by the reach-weighted teacher average within each leaf, so every structure has an attainable floor $F_{\min}$ that no leaf solve on that grouping can beat, and the sweep is read most sharply as the share of the gap $F(0) - F_{\min}$ it takes (Table 12). The unregularized leaf solve sits 11.7 to 27.1 above the floor, and the first grid step captures 80 to 85 percent of that gap under the reported 10^{-12} clip (per seed 69 to 87 percent), leaving an excess of 2.0 to 5.5.

Table 12: The attainable fidelity floor and the share of it the sweep takes, Leduc, IQM over 8 seeds. $F_{\min}$ is the closed-form minimum of $F_0 + F_1$ on the same grouping; the unregularized solve at $\eta = 0$ sits 11.7 to 27.1 above it. Captured is $(F(0) - F(0.01))/(F(0) - F_{\min})$ on the IQMs shown; the per-seed ratios span 69 to 87 percent. All four columns of the sweep are computed on the fixed exploitability-grown structure of Figure 3.

B	$F(0)$	$F_{\min}$	$F(0.01)$	excess	captured	$F(0)/F(0.01)$
4	28.10	1.031	6.499	5.468	79.8%	4.32
8	25.72	1.002	5.636	4.633	81.3%	4.56
16	22.53	0.893	4.168	3.275	84.9%	5.41
32	12.51	0.764	2.740	1.976	83.2%	4.57

Both endpoints and the floor are evaluated under the 10^{-12} clip of Appendix L, so we vary that clip. Re-solving the same 64 cells at $\eta = 0$ and 0.01 on the same structures and re-evaluating all three quantities under clip floors $\varepsilon \in \{10^{-12}, 10^{-6}, 10^{-4}, 10^{-2}\}$ (the re-solves reproduce the sweep's NashConv and fidelity sum exactly, $\max|\Delta| = 0$ over all 64 cells) shows which quantities the convention moves (Table 13). Up to $\varepsilon = 10^{-4}$ the budget IQMs of the floor and the regularized endpoint hold: $F_{\min}$ moves by at most 10^{-3}, and $F(0.01)$ by at most 1.8 percent in 30 of the 32 cells. The two exceptions, seeds 0 and 6 at $B = 8$, fall by 25 percent (8.61 to 6.41) and sit outside the interquartile range, which is why the budget IQM does not register them. Meanwhile $F(0)$ falls by a factor of 2.0 to 3.1 across cells, to budget IQMs of 8.97, 8.25, 7.54 and 4.89. The unregularized excess therefore lives in student probabilities below 10^{-4} on actions the teacher plays (the smallest such probability is below 10^{-13} on every seed), whereas re-evaluating the regularized solve under the same floors leaves its excess where it was, so that excess is not carried by the entries a coarser floor touches. The captured share thus depends on the clip (80 to 85 percent at the reported 10^{-12}, 31 to 52 percent at $\varepsilon = 10^{-4}$), while the absolute excess of the regularized solve, 2.0 to 5.5, moves by at most 2.2 percent up to $\varepsilon = 10^{-4}$ in 30 of 32 cells and by 29 percent in the two exceptions above. The floor is also not an artifact of near-degenerate leaves: recomputing $F_{\min}$ with every leaf probability constrained to be at least ε moves it by at most 2.4×10^{-5} on any seed for $\varepsilon \le 10^{-4}$, and by at most 3.4×10^{-3}, half a percent of its value, at $\varepsilon = 10^{-2}$. The far end reaches the floor itself: at $\eta = 1000$ every one of the 32 cells sits within 1.3×10^{-5} of $F_{\min}$, and that is where the frontier becomes expensive. IQM NashConv 4.64, 3.94, 3.07, and 2.33 at the four budgets, 2.3 to 2.9 times the $\eta = 0$ value at the same budget.

The aligned arm has its own frontier. Re-solving its $B = 8$ and $B = 16$ groupings (seeds 2 to 7, grown on the cluster) at $\eta \in \{0, 0.01, 1000\}$ with the same five-start solve, the first grid step captures 78.4 percent of the gap at $B = 8$ (one structure shared by all six seeds with a stored growth state) and 80.1 percent at $B = 16$ (IQM, range 78 to 82), leaving an excess of 5.21 and 4.49 for a NashConv rise of at most 0.025, and $\eta = 1000$ reaches the floor to within 10^{-5} at both budgets. The cheap first step is therefore not an artifact of the exploitability-grown structure the sweep uses.

Table 13: Clip sensitivity of the frontier endpoints, Leduc, IQM over 8 seeds. The same 64 cells are resolved on the same fixed structure as Table 12 and re-evaluated under clip floors $\varepsilon \in \{10^{-12}, 10^{-4}\}$ (the full grid also covers 10^{-6} and 10^{-2}); captured is $(F(0) - F(0.01))/(F(0) - F_{\min})$ on the budget IQMs, the estimand of Table 12. Raising the floor from the reported 10^{-12} to 10^{-4} leaves $F_{\min}$ and $F(0.01)$ essentially unchanged while $F(0)$ falls by a factor of 2.6 to 3.1.

B	ε	$F(0)$	$F(0.01)$	$F_{\min}$	captured
4	10^{-12}	28.10	6.499	1.031	79.8%
4	10^{-4}	8.97	6.498	1.030	31.1%
8	10^{-12}	25.72	5.636	1.002	81.3%
8	10^{-4}	8.25	5.635	1.001	36.1%
16	10^{-12}	22.53	4.168	0.893	84.9%
16	10^{-4}	7.54	4.154	0.892	50.9%
32	10^{-12}	12.51	2.740	0.764	83.2%
32	10^{-4}	4.89	2.736	0.763	52.2%