SOLVER-GUIDED REASONING FOR MIXED-EQUILIBRIUM STRATEGIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning in large language models (LLMs) is often grounded in human text, human demonstrations, and human-generated rationales. For equilibrium reasoning in complex games, however, relying on human data can be suboptimal. Human play is often guided by intuition and heuristics, while human explanations emphasize individual actions rather than mixed strategies. To strengthen LLM reasoning in games, we study how to articulate equilibrium strategies using solver output. We propose a Mixed-Strategy Decision Tree (MDT), which represents solver mixtures through locally sparse, inspectable decision rules. We evaluate this representation through fixed-LLM strategy prediction, using Scenario-Constrained Counterfactual Sampling (SCCS) to construct same-context hand comparisons. We instantiate the study on No-Limit Texas Hold'em using over **250 million solver-labeled decisions**. With identical reference policies, MDT routing reduces interpolation error by **9.8%** relative to weighting by input distance on held-out boards. In-context comparisons across eight LLM configurations show improved prediction from solver-backed references. With forward-verified inputs and all requests counted, error falls by **37.7%** for DeepSeek and **41.6%** for GPT on policy-filtered targets; independently fixed, training-held-out queries yield reductions of **26.2%** and **20.8%**. Complete River policies and Liar's Dice extend the evaluation to strategic compression and a second game. The results establish useful policy relations in MDT routing and a practical interface for studying how LLMs use solver-backed evidence.

1 INTRODUCTION

Large language models are trained and evaluated with extensive human data, including demonstrations, solutions, rationales, and reasoning traces (Wei et al., 2022; Achiam et al., 2023). For reasoning in complex games, however, human data has important limitations. Play and commentary are selective, and successful human strategies may exploit weaker opponents rather than approximate equilibrium. Even strong human players can underperform solver-based systems (Brown & Sandholm, 2019b; Silver et al., 2018; 2017). Human explanations also tend to frame a decision as choosing one best action. In imperfect-information games, the desired object is often instead a *mixed strategy*: a distribution whose action frequencies help prevent profitable counter-strategies.

Recent poker evaluations make this gap concrete. PokerBench finds that LLMs struggle to recover solver-recommended actions even after task-specific fine-tuning (Zhuang et al., 2025), while the GTO Wizard Benchmark identifies both basic card-representation errors and strategic frequency errors (Provost et al., 2026). An LLM can discuss pot odds, blockers, and equity yet miss how these quantities jointly determine a hand's role in the full range. Moreover, predicting the dominant action or avoiding an immediate value mistake does not establish fidelity to the solver's complete mixture.

The underlying objectives differ. LLMs learn linguistic prediction and reasoning, whereas game solvers compute policies using regret minimization, search, and belief-dependent continuation values (Zinkevich et al., 2007; Moravčík et al., 2017; Brown & Sandholm, 2018). ToolPoker grounds LLM actions and generated reasoning in external solver outputs (Lin et al., 2026). We ask how these numerical targets can be expressed as inspectable mixed-policy computations and comparisons between hands. We call this task *solver articulation*:

How can we extract articulate, verifiable reasoning from the silent optimality of solver-generated data?

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.

We study this question at No-Limit Texas Hold'em (NLH) postflop decision points with two players remaining. Our Mixed-Strategy Decision Tree (MDT) converts public context and solver-derived range- and hand-level summaries into an inspectable policy representation. A mixed strategy combines the action distributions of probabilistically reached leaves; pure-action leaves are a special case. The final model restricts every local router to at most five active summaries. This makes the inputs used at each branching decision explicit and provides a computational basis for inspecting how action mass is distributed.

We evaluate the representation through in-context strategy prediction. Scenario-Constrained Counterfactual Sampling (SCCS) constructs feasible shadow hands that share the public context but differ in policy and MDT route. Their summaries, known policies, and routing differences expose how action frequencies change across private hands. These contrasts form an evaluation input: a fixed LLM receives the comparisons and predicts a hand absent from the references. Direct prompting supplies a baseline without this reference evidence. Input ablations separately examine the effects of reference selection and path displays. Solver outputs construct and score the evaluation, while the target policy remains hidden from the predictor.

The experiments measure both the information retained by MDT and its use in prediction. A three-seed public-board holdout establishes the fidelity–sparsity trade-off: training with sparse support substantially outperforms imposing the same mask afterward. With reference policies fixed, routing lowers interpolation error by 9.8% compared with weighting by input distance. The eight-configuration LLM study measures the prediction benefit of solver-backed references. Two-model tests examine this benefit under matched input definitions on the selected targets and on independently fixed, training-held-out queries. Complete River endgames measure the strategic quality of compressed policies, and Liar's Dice examines the approach in a second game.

This interface connects mixed policies to explicit computations and comparisons that another predictor can inspect. Policy fidelity, routing information, and downstream prediction provide complementary measurements of that connection. Learning from solver-backed experience can complement human demonstrations (Silver & Sutton, 2025). Our work represents an effort toward realizing this new era of experiences.

2 Related Work

Solver-based game reasoning. CFR, subgame solving, and self-play search have enabled strong imperfect-information game agents (Zinkevich et al., 2007; Brown & Sandholm, 2018; 2019b; Moravčík et al., 2017; Brown et al., 2020). Their outputs are primarily numerical prescriptions. We study a communication layer that makes these prescriptions accessible through sparse policies and reference comparisons.

Interpretable policy distillation. Prior work extracts tree or program policies and distills networks into soft decision trees (Bastani et al., 2018; Verma et al., 2018; Frosst & Hinton, 2017). Concept probing also identifies human-understandable structure in AlphaZero (McGrath et al., 2022). We build on policy distillation to study mixed strategies: how local input restrictions affect policy fidelity, how much routing must be displayed to reconstruct a mixture, and what policy information routing retains independently of the action head. Abstraction and hand bucketing also group strategically similar private states (Ganzfried & Sandholm, 2014; Li & Huang, 2025), usually to make solving tractable. Our objective is to expose the small set of summaries used at each routing decision and relate them to action frequencies. Same-context comparisons then provide one way to evaluate how an LLM uses this representation.

LLMs and poker reasoning. Poker benchmarks examine action prediction and opponent-specific play (Gupta, 2023; Zhuang et al., 2025; Provost et al., 2026). We evaluate recovery of a hidden action distribution after receiving solver-derived evidence, with the LLM fixed throughout. Appendix A.4 discusses further distinctions.

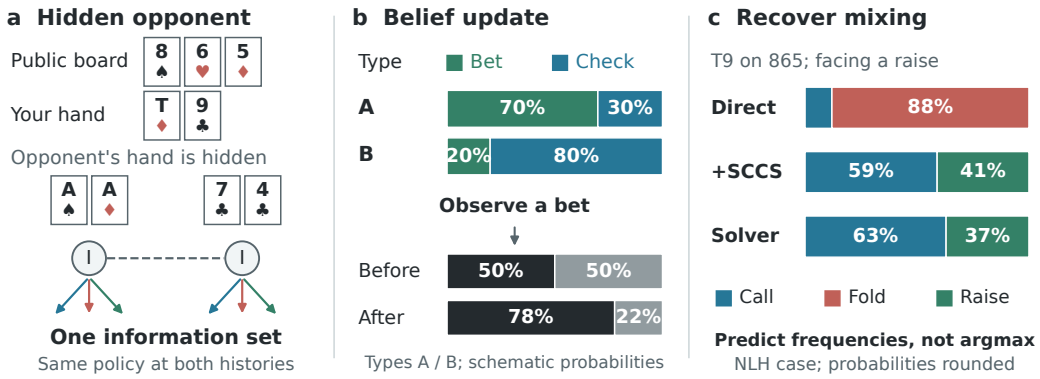

Figure 1: **Explaining a mixed strategy.** (a) The acting player observes its own hand and public state; the opponent's hand remains hidden. (b) Action frequencies change the range observed after an action (schematic probabilities). (c) In this evaluation example, the LLM moves from over-folding to a Call/Raise mixture after receiving same-context references. Bars use rounded probabilities; Raise corresponds to the generic Bet label at this node. Appendix F.7 gives scores and context.

3 Preliminaries

Hidden information and mixed policies. In a two-player zero-sum extensive-form game, an information set $I_i(s, c_i)$ comprises histories sharing the public state s and player i's private hand c_i, but potentially different opponent hands. A behavioral policy assigns a distribution $\pi_i(\cdot \mid I_i) \in \Delta(\mathcal{A}(I_i))$. The public belief state $\beta_s(c_1, c_2)$ describes the hidden private-card assignments consistent with public play; its marginals are the players' *ranges*. In NLH, the public state includes the board, betting history, pot, and stacks. The flop, turn, and river reveal three, four, and five public cards, respectively.

An approximate Nash solver returns a policy π^* and action values Q^*. At information sets reached with positive probability under an exact equilibrium, supported actions have equal value against equilibrium continuations (Nash, 1951; Kuhn, 1953). Consequently, a near-pure prediction can have little local EV loss while getting the mixture wrong. Action frequencies also affect beliefs and subsequent continuation (Brown & Sandholm, 2017; Brown et al., 2020). Figure 1 illustrates this distinction. Formal background and additional motivation are in Appendices A.1 and A.2; poker terminology is in Appendix A.3.

Let h denote a decision with public context, private hand, solver policy $\pi^*(\cdot \mid h)$, and strategic summaries $\mathbf{x}$. MDT approximates the mapping from $\mathbf{x}$ to the mixed policy. Its computation can then be exposed as evidence to another predictor. In the communication evaluation, an LLM receives the target context and permitted evidence and predicts $\tilde{\pi}$. We measure the error of this distribution against the hidden solver policy; this operational measure does not directly score the correctness of a verbal explanation.

4 Mixed-Strategy Decision Trees

The MDT input $\mathbf{x}$ combines public context with range-level and hand-level continuation summaries. In the NLH corpus, the 156 summaries include equity (EQ), expected value (EV), and action-gap quantities under available lines. The solver supplies these numerical inputs; the downstream LLM uses their named values as evidence. Figure 2 separates this input, policy supervision, and rule communication.

Each leaf $l \in \mathcal{L}$ stores a distribution q_l over legal actions. At internal node n, a local router computes branch probabilities $p_{n,c}(\mathbf{x})$. A leaf receives the product of the branch probabilities on its path, and the policy mixes leaf distributions:

$$\rho_l(\mathbf{x}) = \prod_{(n,c)\in \mathrm{Path}(l)} p_{n,c}(\mathbf{x}), \qquad \hat{\pi}(a \mid \mathbf{x}) = \sum_{l\in\mathcal{L}} \rho_l(\mathbf{x}) q_l(a \mid \mathbf{x}). \tag{1}$$

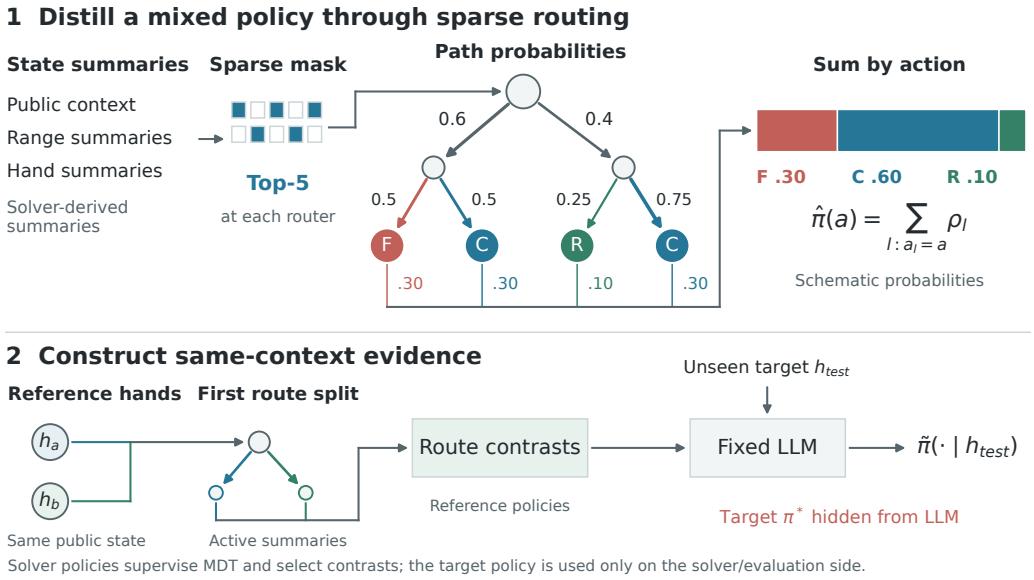

Figure 2: **Representing and communicating mixed policies.** Top: a schematic pure-leaf fragment illustrates how equal-action masses form a mixture; NLH communication also permits soft leaves. Bottom: SCCS constructs the same-context reference condition for LLM evaluation, displaying reference policies, numerical summaries and routing contrasts. The target policy is hidden from the predictor. Reference-first tests use target labels for selection; target-first tests fix queries without consulting their labels.

The final router uses a node-local mask $\mathbf{m}_n$ with at most $K = 5$ active summaries:

$$\mathbf{p}_n(\mathbf{x}) = \mathrm{Normalize}\big(r_n(\mathbf{x} \odot \mathbf{m}_n)\big), \qquad \|\mathbf{m}_n\|_0 \le 5. \tag{2}$$

Hard MDT selects summaries discretely and routes probabilistically. Normalized branch and leaf distributions make Equation 1 a valid policy. Pure leaves have $q_l(a) = \mathbf{1}[a_l = a]$: two Call leaves of mass 0.30 contribute 0.60 together (Figure 2). NLH communication uses soft leaves and normalized softplus routing; River deployments use pure leaves (Appendix B.1). The policy aggregates all leaves, including those outside a displayed route.

The router can use a sparse linear logit map or an additive scalar-to-logit map:

$$r_n(\mathbf{x} \odot \mathbf{m}_n) = W_n(\mathbf{x} \odot \mathbf{m}_n) + \mathbf{b}_n \quad \text{or} \quad \sum_{j:m_{n,j}=1} \mathbf{h}_{n,j}(x_j). \tag{3}$$

Both parameterizations expose at most five summaries *per router*. The selected inputs and their numerical transformations specify each branch computation; path probabilities and leaf distributions then reconstruct the mixed policy. Routers may select different input sets, so the complete model can use more than five (Appendix B.1).

Training combines reach-weighted Smooth-L_1 policy fit, nonnegative one-step EV shortfall and a target-probability hinge, weighted 50, 10 and 2 (Appendix B.2). Communication checkpoints use validation L_1. EV shortfall uses fixed continuations; River best-response evaluation is separate (Section 5.3).

Soft sparsity adds weight sparsity, routing-entropy, and orthogonality regularizers to differentiable routing; weak nonzero inputs can remain outside the displayed summaries. Hard distillation restricts the inputs used in the forward computation (Appendix B.2).

Hard distillation fits a globally gated ResNet using 50 summaries, a tree teacher with two-layer MLP routers, then a student with straight-through Top-5 masks (Appendix B.2). The student uses only selected summaries in its forward computation.

Because the strategy combines several leaves, a displayed route need not represent its full mixture. We can display the shortest descending-weight set of paths retaining a chosen mass m, to-

Table 1: **Policy fidelity on held-out public boards.** Mean and sample SD over three training seeds on the same 10,240 hands. All students have 172,951 parameters; Top-50→5 changes only the inference mask. Lower summed action L_1 is better.

Model	Summed L_1	Seed SD
Global-gated ResNet	0.2804	0.0016
Tree teacher	0.3974	0.0038
Trained Top-5 MDT	0.4251	0.0024
Trained Top-10 MDT	0.3889	0.0030
Trained Top-50 MDT	0.3549	0.0037
Top-50→5 post-hoc mask	0.7395	0.0202

gether with the omitted mass and active inputs. The normalized displayed mixture p_S satisfies $\|p - p_S\|_1 \leq 2(1 - m)$. This gives a numerical approximation budget alongside the number of paths and inputs to inspect (Appendices D.2 and D.3).

Routing also relates different hands through their distribution of mass over shared leaves. For hands h and h_j, their overlap is

$$\kappa(h, h_j) = \sum_{l \in \mathcal{L}} \min\{\rho_l(\mathbf{x}_h), \rho_l(\mathbf{x}_{h_j})\}. \tag{4}$$

This statistic uses all 81 paths of the communication tree and ranges from zero for disjoint routing to one for identical routing. It provides a learned relation that can be used to weight reference policies or supplied as evidence to an LLM. These two uses are evaluated separately below.

5 Experiments and Results

We query a commercial high-end NLH solver for over 250 million postflop decisions: approximately 16M flop and 235M turn samples, with mixed policies and continuation quantities. The configuration is six-player, 100-big-blind, no-rake NLH, evaluated at postflop spots with two players remaining. Five preflop scenarios cover SB vs BB single-raised/3-bet pots, BTN vs BB single-raised/3-bet pots, and BTN vs SB 3-bet pots. Data spans 1,755 flop textures; representative betting branches are extended by five sampled turn cards. Response sizes follow the solver abstraction, with aggressive/all-in actions mapped to its reported generic action categories. These corpus samples are distinct from the complete River trees evaluated below.

The experiments examine the representation itself, its use in LLM strategy prediction, and the strategic quality of complete compressed policies. The main error metric is the L_1 distance between action distributions. NLH results use summed L_1 unless marked action-averaged, which divides by the three abstract action slots, including masked zero slots. The Liar's Dice communication metric instead averages over legal actions. Lower values indicate closer agreement with the solver. Each comparison uses a common population and metric; the different evaluation populations are reported separately.

5.1 Policy Fidelity and Sparse Routing

We first examine how well MDT preserves a solver policy when each router can use only a few summaries. Models are fitted on 256 training and 64 validation board groups and evaluated on 64 disjoint test groups. Boards equivalent under suit permutations stay together across all scenarios, seats and histories. Ten scenario/seat subsets each supply one node per group and sixteen uniformly sampled positive-reach hands per node. The resulting test set contains 10,240 hands across 640 nodes, without selection by policy distance or mixing. Inputs retain the 156 oracle-derived summaries.

Within each seed, students share initialization, teachers, topology, data, losses and update count; only local support changes. The three seeds share the board partition. Validation selects checkpoints under their native masks. Table 1 shows a fidelity cost as local support decreases. For the primary seed, Top-5 exceeds Top-50 error by 0.0656, with seven-comparison-adjusted 99.2857% board-cluster interval $[0.0580, 0.0731]$. Training with sparse support is substantially better than removing inputs afterward: error falls from 0.7202 for the post-hoc mask to 0.4247 for trained Top-5, a difference of 0.2956 $[0.2757, 0.3146]$. Both other seeds give the same ordering (Appendix C.1).

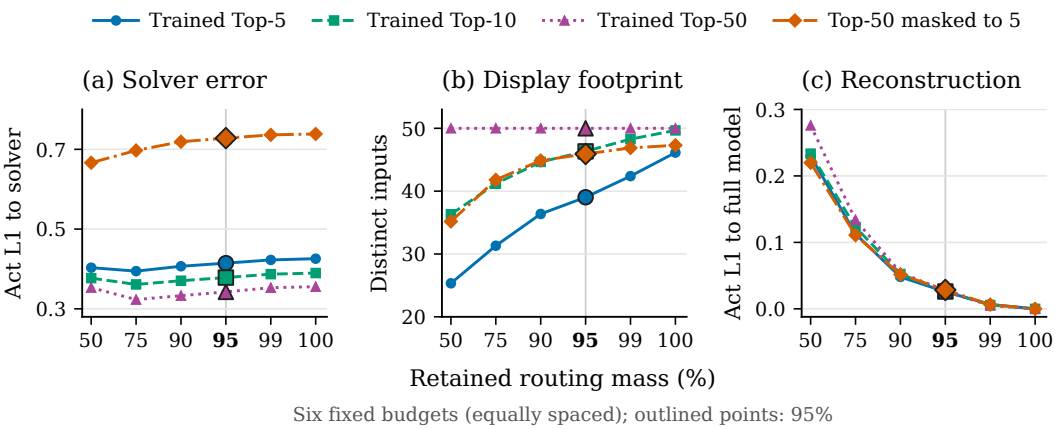

Figure 3: **Policy fidelity and display size on held-out boards.** All six retained-mass budgets are shown at equal spacing; outlined points retain 95%. Curves average three seeds on the same 7,680 queries per seed. Panels show solver error, distinct inputs across displayed routers, and error relative to each model's complete policy. Post-hoc masking restricts a Top-50-trained model to five inputs per router.

Table 2: **Policy information in routing.** Fixed-reference numerical interpolation on 7,680 held-out queries per seed, averaged over three seeds. Every rule uses the same four reference labels. Lower summed L_1 is better.

Reference weighting	Summed L_1
Raw input distance	0.6008
Standardized input distance	0.6002
Soft-routing overlap	0.5416
Permuted overlap assignment	0.7155

Local input count alone does not determine the size of a complete display. At 95% retained mass, Top-5 uses 16.60 leaves and 39.03 distinct summaries on average, with 0.0258 discrepancy from the complete MDT. Top-50 uses all 50 summaries. These quantities describe what must be inspected to reconstruct the mixture at a specified fidelity; they complement the five-input bound at each router. Appendix D.3 reports the same measurements on the communication checkpoints.

Figure 3 separates local input support during training from retained routing mass at display time. The former changes the learned policy; the latter approximates a fixed policy. Both solver error and reconstruction error are measured at six preselected mass budgets, since a shorter display can occasionally approach the solver while moving away from its own full model.

We next ask whether the learned routing captures relations between hands that are useful independently of the action head. At each held-out node, four randomly ordered hands supply reference labels and the remaining twelve are queries. We weight the same four labels by input distance or routing overlap (Equation 4). A cyclic permutation of overlap weights provides an assignment control. No target policy, target model prediction or leaf strategy enters these weights.

With identical reference policies, MDT routing lowers interpolation error by 9.8% relative to standardized input distance (Table 2). The primary-seed gain is 0.0584, with adjusted interval $[0.0456, 0.0715]$; correct assignment improves over permutation by 0.1751 $[0.1493, 0.2038]$. Both gains recur in each seed. MDT learns how the input summaries jointly distribute probability mass over shared leaves. With inputs and reference labels fixed, the comparison measures the predictive value of this learned relation. The complete MDT action head, which achieves 0.4258 on these queries, is withheld. This numerical result motivates the next evaluation: presenting solver evidence and MDT-derived comparisons to a fixed LLM for strategy prediction.

Table 3: **Cross-model NLH communication on reference-first targets.** Action-averaged L_1 to the hidden solver policy, lower is better. Means use each condition's successfully parsed outputs; coverage differs across conditions and configurations. The last row weights configurations equally. Counts, standard errors and input/execution differences are in Appendices F.2 and F.3.

LLM/API configuration	Direct	+Summaries	Greedy Route-only	SCCS
Gemini-3.1 Flash	.211	.223	.144	.099
Gemini-3.1 Pro low	.235	.299	.228	.106
Gemini-3.1 Pro high	.268	.303	.205	.090
DeepSeek-V4 Flash	.185	.263	.212	.116
DeepSeek-V4 Pro	.254	.255	.174	.112
GPT-5.4	.187	.216	.124	.095
GPT-5.5 low	.158	.241	.150	.096
GPT-5.5 high	.186	.248	.147	.086
Mean	.211	.256	.173	**.100**

5.2 In-Context Strategy Prediction

We evaluate MDT-supported strategy communication through SCCS-constructed reference comparisons. A fixed LLM receives other hands' policies, numerical summaries and MDT routing contrasts within the same public context, and predicts the target hand's solver strategy. In this reference condition, the target's solver policy, MDT prediction and route remain hidden. Providing the target's complete MDT computation would make its model prediction directly recoverable. SCCS instead tests prediction from evidence about reference hands.

Direct provides the state and target hand; **Direct+Summaries** adds target numerical inputs. **Greedy Route-only** adds that target's own greedy MDT path, without labeled references. These target-side baselines complement the reference-based task. We report the complete reference condition first, followed by input ablations.

To construct comparisons between different mixed strategies, SCCS selects three feasible shadows of a reference hand by MDT-policy difference and first route divergence, holding board, betting history and acting position fixed. Minimum summed policy distances are 0.05 from the anchor and 0.03 between shadows (Algorithm 1). These contrasts include hands sharing the same most likely action. A separate target is withheld from the four references.

Reference-first sampling selects 100 targets using solver-distance and mixing filters after constructing references, retaining 100 of 983 attempts (Appendix E.3). Target-first sampling fixes queries before constructing evidence, without consulting their solver policies. Both hide the target policy from the LLM. Table 3 shows the lowest reported error in the reference condition for all eight configurations. Its equal-configuration mean falls from 0.211 to 0.100 relative to Direct (52.6%) and from 0.173 to 0.100 relative to Greedy Route-only (42.2%), computed before rounding.

On jointly parsed Greedy Route-only and SCCS targets, error falls by 41.2%: paired difference -0.0701, target-bootstrap 95% interval $[-0.0938, -0.0460]$. Rendering, numerical-input, execution and coverage differences make this a package-level comparison (Appendices F.3 and E.4).

With forward-verified inputs on the same 100 targets, Direct-relative gains are 37.7% for DeepSeek-V4 Pro and 41.6% for GPT (`gpt-6-sol`, medium reasoning). Each condition uses one attempt; failures receive a legal-uniform prediction. Adjusted comparisons favor references over all three baselines (Table 4; Appendix F.6), including Direct sensitivity checks (Appendix I.3).

We also fix 320 queries from 64 board groups excluded from MDT training and validation, without consulting target policies. Each of ten scenario/seat subsets contributes 32 queries; all remain, including 22 predefined uniform-reference fallbacks. DeepSeek-Flash and the same GPT configuration improve over Direct and Direct+Summaries, with Direct-relative reductions of 26.2% and 20.8%. Four prespecified contrasts per model use 50,000 paired board-cluster resamples and 98.75% intervals for Bonferroni-adjusted 95% family coverage (Appendix F.1).

Table 4: **Reference evidence improves strategy prediction.** Summed action L_1, including invalid-output fallback. DeepSeek uses V4 Pro for the 100-target test and Flash for the 320-query test; GPT is unchanged. Results and coverage are reported separately by population (Appendices F.6 and F.1).

Input	DeepSeek	GPT
Reference-first targets ($N = 100$)		
Direct	0.5163	0.5064
SCCS references	**0.3216**	**0.2960**
Independently fixed, training-held-out queries ($N = 320$)		
Direct	0.8538	0.5093
Direct+Summaries	0.7646	0.4928
SCCS references	**0.6305**	**0.4035**

These results establish a benefit from the complete reference evidence across model configurations and on both policy-filtered and training-held-out queries. The following ablations examine individual inputs.

Input ablations. On the 100 reference-first targets, Summaries and Greedy Route-only have identical numerical inputs. The latter adds the *target hand's own* locally highest-probability path, node contributions, branch probabilities and leaf prototype. Both predict that hand's full solver strategy without labeled references.

Table 5: **Target-path input ablation.** Summed action L_1 on all 100 targets with identical summaries and legal-uniform fallback. Greedy Route-only adds a single path. Counts precede fallback; DeepSeek uses V4 Pro.

Input	DeepSeek L_1	GPT L_1	DS valid	GPT valid
Direct+Summaries	1.0111	0.4838	99	100
Greedy Route-only	0.9774	0.4735	100	100

The full MDT mixes all leaves using products of node-level branch probabilities (Equation 1). The displayed greedy leaf carries only 20.7% of total mass on average (Appendix D.1). This target-trace baseline tests partial computational evidence, separately from SCCS reference-based prediction. Error reductions are 3.3% and 2.1%, with paired intervals containing zero (Table 5; Appendix F.6).

The reference-selection ablation compares SCCS with uniform and query-nearest references. On independently fixed queries, SCCS does not significantly outperform uniform selection, and query-nearest references are more accurate (Appendix F.1). SCCS constructs policy contrasts around an anchor; nearest selection retrieves similar inputs for the target.

With the reference set fixed, a separate ablation tests the added path display. All conditions retain the same reference identities, solver labels, summaries and reference MDT predictions; only the *reference hands'* path text varies. On 300 target-first queries, DeepSeek favors faithful over shuffled paths, but neither model establishes a gain over common evidence. On 320 training-held-out queries, adjusted contrasts against common evidence, shuffling and repetition include zero (Appendix G.1).

5.3 Strategic Quality and a Second Game

We evaluate strategic compression on two complete River trees from the published root states of Subgames 3 and 4 in Brown & Sandholm (2019a), retaining their root ranges and using our two-size betting abstraction through fold or showdown. Best-response evaluation complements local L_1 and fixed-continuation EV gap. Soft and Hard MDT share a depth-4 ternary architecture with pure-action leaves; Hard uses at most five summaries per router. Inputs exclude action Q/value quantities, target probabilities, and exact private-card indicators.

One MDT per player is shared across that player's decision states, and strategic evaluation traverses its complete policy within the River game. High-precision local DCFR targets have residual exploitabilities of 0.009034 and 0.009907 milli-big-blinds per game (mbb/g), respectively. For a deployed

Table 6: **Exact River exploitability, normalized by root pot.** Hard–Soft is in percentage points (pp). Deep CFR is the same-tree average strategy at iteration 800, not a converged bound. Appendix J.2 retains absolute mbb/g and other diagnostics.

River state	Soft MDT	Hard Top-5	Hard–Soft	Deep CFR
Subgame 3 (5 bb)	0.1338%	0.3435%	0.2096 pp	0.8615%
Subgame 4 (37.5 bb)	0.1467%	0.3548%	0.2081 pp	0.3560%

profile $\hat{\pi}$, we compute

$$\text{Exploitability}(\hat{\pi}) = \frac{1}{2}\sum_{i=1}^{2}\left[\max_{\pi_i'} u_i(\pi_i', \hat{\pi}_{-i}) - u_i(\hat{\pi})\right]. \tag{5}$$

This recomputes the opponent's best response to the entire policy. Soft and Hard share architecture and evaluation trees, but use different optimization schedules; their comparison measures realized deployment costs (Appendix J.2).

Hard MDT incurs exploitability below 0.36% of the root pot, approximately 0.21 percentage points above Soft MDT (Table 6). Subgame 4's root pot is 7.5 times larger. Deep CFR (Brown et al., 2019) uses 10,000 traversals per player per iteration as a finite-training reference. Training and selection use these same trees, evaluating *in-domain strategic compression*.

Liar's Dice tests the representation and communication evaluation outside NLH. The two-player game uses two three-faced dice per player and bids up to quantity three, with targets from a full-tree CFR+ solver. Six-fold cross-fitting holds out one of six unordered private dice types per public context. MDT uses 17 public-history, posterior-belief, private-dice, and equity summaries, excluding per-action continuation values and Q-gaps from both inputs and prompts.

The exploratory communication evaluation uses 31 held-out targets and the four NLH input conditions. Averaging the parsed-output means across eight configurations, legal-action-averaged solver L_1 is 0.174 for Direct, 0.171 with summaries, 0.116 for Greedy Route-only and 0.105 for SCCS references. The margin over Greedy Route-only is smaller and heterogeneous across configurations. Appendix K.1 reports all per-configuration results, coverage and evaluation details.

Assembling the six held-out folds covers every information set. Against the fixed approximate-Nash opponent, full-policy value loss averages 0.058 per game over the two seats on the $[-1, +1]$ scale. This fixed-opponent metric complements adaptive River exploitability.

6 Conclusion

MDT represents solver mixtures through locally sparse routing, making the computation of an action distribution explicit. Matched fitting measures the fidelity cost of local sparsity, and routing improves reference interpolation on held-out boards with labels fixed. In-context comparisons show that solver-backed examples improve LLM strategy prediction across model configurations and independently fixed queries. These communication gains concern the complete reference evidence. Complete River policies and Liar's Dice connect the representation to strategic compression and a second game. Together, these results provide a concrete basis for studying how numerical solver knowledge can be represented and communicated.

AI USE STATEMENT

Generative language models interpret structured solver evidence and predict strategies in the reported experiments; illustrative examples also contain generated verbal explanations. AI assistance was used for manuscript editing, checking original records, diagrams, experiment design, data preparation, implementation, and statistical analysis. We take responsibility for the manuscript, code, figures, and our main claims.

ETHICS STATEMENT

This work studies solver-assisted analysis of imperfect-information games, not advice for real-money gambling. Solver licensing and data-redistribution restrictions must be respected. The reported language-model outputs can contain factual or strategic errors and should not be treated as certified instructions for play.

REPRODUCIBILITY STATEMENT

The router, optimization curriculum, SCCS algorithm, metric definitions, evaluation coverage, and prompt templates are specified in the appendices. The commercial NLH corpus requires licensed oracle access; the Liar's Dice game and solver offer a reproducible counterpart. The River appendix identifies public starting states, local tree construction, training/selection conditions, and the finite-training Deep CFR comparison. These descriptions distinguish releasable components from the licensed corpus. The supplementary implementations provide retained-mass displays and reference weighting, with synthetic tests that require no private data or model weights. They verify mixture reconstruction and interpolation, not downstream reasoning (Appendices D.2 and C.1).

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

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. doi: 10.1609/aaai.v33i01.33011829.

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 *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 793–802. PMLR, 2019. URL `https://proceedings.mlr.press/v97/brown19b.html`.

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.

Nicholas Frosst and Geoffrey Hinton. Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*, 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.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

Jiaxian Guo, Bo Yang, Paul Yoo, Bill Yuchen Lin, Yusuke Iwasawa, and Yutaka Matsuo. Suspicion-agent: Playing imperfect information games with theory of mind aware gpt-4. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=F2yGbwXJAi.

Akshat Gupta. Are chatgpt and gpt-4 good poker players? a pre-flop analysis. *arXiv preprint arXiv:2308.12466*, 2023.

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

Chenghao Huang, Yanbo Cao, Yinlong Wen, Tao Zhou, and Yanru Zhang. Pokergpt: An end-to-end lightweight solver for multi-player texas hold'em via large language model. *arXiv preprint arXiv:2401.06781*, 2024.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pp. 5338–5348, 2020.

Harold W. Kuhn. Extensive games and the problem of information. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games II*, volume 28 of *Annals of Mathematics Studies*, pp. 193–216. Princeton University Press, 1953.

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.

Minhua Lin, Enyan Dai, Hui Liu, Xianfeng Tang, Yuliang Yan, Zhenwei Dai, Jingying Zeng, Zhiwei Zhang, Fali Wang, Hongcheng Gao, Chen Luo, Xiang Zhang, Qi He, and Suhang Wang. How far are llms from professional poker players? revisiting game-theoretic reasoning with agentic tool use. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=vV54ShHvGi.

Thomas McGrath, Andrei Kapishnikov, Nenad Tomašev, Adam Pearce, Martin Wattenberg, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. Acquisition of chess knowledge in alphazero. *Proceedings of the National Academy of Sciences*, 119(47):e2206625119, 2022.

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, 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.

John Nash. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, 1951.

Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*. MIT Press, 1994.

Marc-Antoine Provost, Nejc Ilenic, Christopher Solinas, and Philippe Beardsell. Gto wizard benchmark, 2026. URL https://arxiv.org/abs/2603.23660.

Andrei A Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

David Silver and Richard S Sutton. Welcome to the era of experience. Preprint, 2025. URL https://storage.googleapis.com/deepmind-media/Era-of-Experience%20/The%20Era%20of%20Experience%20Paper.pdf.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419): 1140–1144, 2018. doi: 10.1126/science.aar6404.

Abhinav Verma, Vijayaraghavan Murali, Rishabh Singh, Pushmeet Kohli, and Swarat Chaudhuri. Programmatically interpretable reinforcement learning. In *International Conference on Machine Learning*, pp. 5045–5054, 2018.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Richard Zhuang, Akshat Gupta, Richard Yang, Aniket Rahane, Zhengyu Li, and Gopala Anumanchipalli. Pokerbench: Training large language models to become professional poker players. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26175–26182, 2025.

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 Game Background and Evaluation Concepts

A.1 Mixed Policies and Public Beliefs

Action-conditioned ranges. To see the coupling, write $b_{i,s}(c_i)$ for the acting player's range at public state s. Observing action a, before any further chance event, changes the public posterior to

$$b_{i,s,a}(c_i) = \frac{b_{i,s}(c_i)\,\pi_i(a \mid I_i(s,c_i))}{\sum_{c_i'} b_{i,s}(c_i')\,\pi_i(a \mid I_i(s,c_i'))}, \tag{6}$$

when the denominator is positive. A hand's frequency therefore changes not only its immediate action but also the composition of the range taking that action. The schematic example in Figure 1b starts with equally likely private types that bet with probabilities 0.7 and 0.2; after a bet, their posterior shares are $7/9$ and $2/9$. This is why selecting a plausible action for one hand is insufficient to explain a range-coupled mixture.

Imperfect-information extensive-form games. We use the two-player zero-sum game $\mathcal{G} = (\mathcal{N}, \mathcal{H}, \mathcal{Z}, \mathcal{A}, P, u, \mathcal{I})$: $\mathcal{N} = \{1, 2\}$, histories $\mathcal{H}$, terminal histories $\mathcal{Z}$, legal actions $\mathcal{A}(h)$, player/chance function P, payoffs satisfying $u_1(z) + u_2(z) = 0$, and information sets $\mathcal{I}_i$. Histories in one information set share public events and player i's private hand but may differ in the opponent's hand. A behavioral strategy assigns $\pi_i(\cdot \mid I) \in \Delta(\mathcal{A}(I))$. We reserve h for a decision history and z for a terminal history. Appendix A.3 defines the poker vocabulary.

Nash equilibrium and GTO strategies. A strategy profile $\pi = (\pi_1, \pi_2)$ induces an expected utility $u_i(\pi)$ by integrating terminal utilities over chance outcomes and both players' randomized actions. A Nash equilibrium (NE) is a profile π^* such that no player can improve by unilateral deviation (Nash, 1951):

$$u_i(\pi_i^*, \pi_{-i}^*) \geq u_i(\pi_i', \pi_{-i}^*) \qquad \forall i \in \mathcal{N},\ \forall \pi_i'. \tag{7}$$

In poker terminology, a Game-Theoretic Optimal (GTO) strategy is an approximate NE strategy computed by a solver. Modern poker solvers are commonly based on regret minimization, search, public-belief reasoning, and subgame solving (Zinkevich et al., 2007; Bowling et al., 2015; Moravčík et al., 2017; Brown & Sandholm, 2018). Solver outputs are commonly reported as mixed action distributions together with action values or counterfactual values. For a decision history h, $\pi^*(\cdot \mid h)$ denotes the computed action distribution, and $Q^*(h, a)$ denotes the value associated with legal action $a \in \mathcal{A}(h)$. The notation $Q^*(I, a)$ is used when discussing the corresponding information-set formalism.

Public belief states and ranges. At public state s, the public belief state is $\beta_s(c_1, c_2) = \Pr(c_1, c_2 \mid s)$; a player's range is its marginal (Brown et al., 2020). The solver maps public context, private hand and belief to $\pi_i^*(\cdot \mid I_i(s, c_i))$. Continuation values and both players' ranges can therefore affect one hand's action frequencies.

Indifference and mixed strategies in Nash equilibrium. Randomization controls information revealed by actions. A local policy changes the range seen by the opponent, coupling decisions across private hands.

At an information set reached with positive probability under an exact equilibrium, shifting probability from a lower-value supported action to a higher-value action would create a profitable unilateral deviation. Supported actions therefore have equal value against equilibrium continuations, and unsupported actions have no greater value (Nash, 1951; Osborne & Rubinstein, 1994). In games with perfect recall, behavioral strategies represent a player's randomized choices at information sets (Kuhn, 1953), and CFR-style algorithms approach equilibrium by minimizing counterfactual regret (Zinkevich et al., 2007). For such a reached information set I,

$$Q_i^*(I, a) = V_i^*(I) \qquad \text{for } a \text{ with } \pi_i^*(a \mid I) > 0, \tag{8}$$

with $Q_i^*(I, a) \leq V_i^*(I)$ for actions outside the support. A global approximate-equilibrium tolerance does not generally give the same bound on every conditional action-value difference, particularly at rarely reached information sets. Local solver quantities should be interpreted with their own numerical accuracy.

Global dependence. A reached imperfect-information subgame can depend on strategies in unreached parts of the game, preventing independent solution from the visible state alone (Brown & Sandholm, 2017).

A.2 Range-Level Frequency Prediction

Equilibrium targets specify frequencies across a range, not only a plausible action for one hand. Value hands, bluffs, draws and blocker-related candidates can share an action, while strong hands sometimes check. Belief updates couple these assignments (Equation 6). CFR and search-based poker systems address this coupling numerically (Zinkevich et al., 2007; Moravčík et al., 2017; Brown & Sandholm, 2018; 2019b).

Public poker text usually discusses selected hands or exploitative decisions. PokerBench reports GPT-4 exact-match accuracy of 53.55% before task-specific fine-tuning (Zhuang et al., 2025); earlier evaluations likewise separate verbal familiarity from GTO play (Gupta, 2023). MDT distills the solver's mixed policy into a learned representation. SCCS constructs same-context contrasts and labeled evidence for evaluating that representation with fixed LLMs. Prediction of a hidden policy is the measured outcome; direct semantic comprehension is not measured.

A.3 Game Rules and Vocabulary

A.3.1 Rules of No-Limit Texas Hold'em

The corpus uses six-player, 100BB, no-rake configurations with two players remaining at each evaluated postflop decision. Each player holds two private cards and forms the best five-card hand from them and up to five community cards. The button moves clockwise; the next two seats post the small and big blinds (SB/BB). Preflop action starts left of BB; postflop action starts left of the button. The first remaining actor is out of position (OOP), the second in position (IP).

Preflop precedes community cards; flop, turn and river reveal three, four and five public cards. The corpus focuses on flop and turn. Players may check when no bet is faced, bet, call, fold or raise subject to the current line, up to their remaining stack. The abstract aggressive slot groups bet/raise actions. Bet sizes are expressed relative to the pot (e.g., $0.5\times$ or $1.79\times$); SPR is effective remaining stack divided by pot. Hand order is high card, pair, two pair, three of a kind, straight, flush, full house, four of a kind, straight flush. A kicker breaks otherwise tied made hands; historical feature names need not implement this meaning (Appendix D.4).

A.3.2 Common Poker Terms

Term	Meaning in this paper
Action line / history	The sequence of previous betting actions and public card events leading to the current decision point.
All-in	A bet or raise that commits a player's entire remaining stack.
Air	A hand with little or no current showdown value and limited immediate equity.
Backdoor draw	A draw that needs favorable cards on both later streets to complete, such as needing both turn and river to make a flush.
Bet size	The amount placed into the pot, often normalized by the current pot size.
Blocker	A card in a player's hand that removes combinations from the opponent's possible range, often reducing the chance that the opponent holds strong continuing hands.
Board	The public community cards visible to both players.
Board texture	Strategic properties of the board, such as pairedness, connectedness, straight potential, and flush potential.
Bluff	An aggressive action with a hand that is unlikely to be best if called, used to make better hands fold.
Call	Matching the current bet to continue in the hand.

Term	Meaning in this paper
Check	Passing the action when no bet is faced.
Continuation bet	A postflop bet made by the player who was the previous aggressor, commonly abbreviated as c-bet.
Draw	A hand that is not currently strong but can improve to a strong hand on later community cards.
Equity (EQ)	The probability, or solver-computed share, that a hand or range wins at showdown under the relevant future-card distribution.
Expected value (EV)	The expected payoff of a hand, range, or action under the solver's continuation strategy.
Fold equity	The value gained from the probability that an opponent folds to an aggressive action.
Flush draw	A draw to five cards of the same suit.
Gutshot	An inside straight draw that can complete with one specific rank.
GTO	Game-Theoretic Optimal; in this paper, an approximate Nash-equilibrium poker strategy computed by a solver.
Hand	Usually the player's two private cards, and sometimes the resulting best five-card category depending on context.
In position (IP)	The player who acts second on postflop streets.
MDF	Minimum defense frequency, a pot-odds-derived threshold describing how often a range must continue to avoid being immediately exploitable by a bet.
Mixed strategy	A probability distribution over legal actions at a decision point.
Nuts	The strongest possible hand, or class of strongest hands, for the current board.
Nut advantage	A range-level advantage in the frequency or equity of nut-class hands.
Offsuit	A two-card private hand whose cards have different suits.
Open-ended straight draw (OESD)	A straight draw that can complete with a card on either end of the sequence.
Out of position (OOP)	The player who acts first on postflop streets.
Overcard	A private card higher than every card on the board.
Postflop	Any decision after the flop has been dealt; includes flop, turn, and river.
Pot odds	The price offered by the pot for calling a bet, usually expressed as a required equity threshold.
Protection bet	A bet with a vulnerable made hand or semi-made hand intended to deny equity to hands that can improve.
Range	The probability distribution over private hands a player can hold after conditioning on public cards and betting history.
Range advantage	A range-level equity or EV edge for one player over the other in the current public state.
River	The fifth community card and final betting street.
Semi-bluff	A bluffing bet or raise with a hand that can improve to a strong hand on later streets.
Set	Three of a kind made with a pocket pair and one matching board card.
Showdown value	The ability of a hand to win if betting stops and the hand reaches showdown.
Single-raised pot	A pot where the preflop action contains one raise and no 3-bet.
Suited	A two-card private hand whose cards share the same suit.
Thin value	A value bet with a hand that is ahead of some calling hands but not strong enough to be clearly dominant.
Three-bet pot / 3-bet pot	A pot where the preflop action contains a raise and then a re-raise.
Trap	A passive action with a very strong hand, used to keep weaker hands or bluffs in the opponent's range.
Turn	The fourth community card and the betting street after it is dealt.
Unblocker	A card property that leaves the opponent's folding range relatively intact, which can improve bluff quality in some contexts.

Term	Meaning in this paper
Value bet	A bet made with a hand expected to be called by worse hands often enough to profit.

A.4 Related Policy and Poker Evaluations

Solver policies and interpretable representations. CFR, search and public-belief methods provide numerical equilibrium targets (Zinkevich et al., 2007; Bowling et al., 2015; Brown & Sandholm, 2018; 2019b; Moravčík et al., 2017; Brown et al., 2020). Policy distillation compresses teachers (Hinton et al., 2015; Rusu et al., 2015); concept bottlenecks expose semantic variables (Koh et al., 2020); mixture-of-experts models use input-dependent routing (Jacobs et al., 1991; Shazeer et al., 2017). Tree/program extraction and differentiable trees provide related policy representations (Bastani et al., 2018; Verma et al., 2018; Frosst & Hinton, 2017), while poker abstraction groups strategically similar hands (Ganzfried & Sandholm, 2014; Li & Huang, 2025). Concept analyses of AlphaZero study understandable structure in perfect-information play (McGrath et al., 2022). Our evaluation concerns reference evidence for complete mixed distributions.

LLM poker evaluation. PokerGPT evaluates action prediction, behavioral statistics and Slumbot play; Suspicion-Agent reports finite-match outcomes in Leduc; PokerBench uses dominant-action and exact-bet-size labels (Huang et al., 2024; Guo et al., 2024; Zhuang et al., 2025). The GTO Wizard Benchmark reports its approximate-Nash agent beating Slumbot by 194 ± 41 mbb/hand and GPT-4 losing $1{,}362 \pm 256$ mbb/hand to that agent (Provost et al., 2026). Those HUNL protocols differ from ours and their values are not directly comparable with local L_1. Action accuracy discards supported-action frequencies; fixed-opponent play and complete-policy exploitability also answer different questions from mixture prediction. Work on rationales and synthetic educational data motivates evidence-based learning (Wei et al., 2022; Zelikman et al., 2022; Gunasekar et al., 2023), but our communication tests keep downstream LLM parameters fixed.

B MDT Parameterization and Optimization

B.1 Routers, Leaves and Input Indexing

Each internal node n computes a branch distribution from a small set of active summaries. Let $\mathbf{m}_n \in \{0,1\}^D$ denote the node-local hard mask and let $\|\mathbf{m}_n\|_0 \le K$. The branch logits are

$$\mathbf{g}_n(\mathbf{x}) = r_n(\mathbf{x} \odot \mathbf{m}_n), \qquad \mathbf{p}_n(\mathbf{x}) = \mathrm{Normalize}(\mathbf{g}_n(\mathbf{x})), \tag{9}$$

where r_n is a small local router. In the ablations in Table 14, r_n can be either a sparse linear map,

$$\mathbf{g}_n(\mathbf{x}) = W_n(\mathbf{x} \odot \mathbf{m}_n) + \mathbf{b}_n, \tag{10}$$

or an additive one-dimensional router, implemented as a neural additive model (NAM),

$$\mathbf{g}_n(\mathbf{x}) = \sum_{j:m_{n,j}=1} \mathbf{h}_{n,j}(x_j), \tag{11}$$

with each $\mathbf{h}_{n,j}$ implemented as a small scalar-to-logit network. These choices expose a node-local set of at most K summaries and the resulting branch probabilities.

NLH implementation and display conventions. The NLH communication model uses normalized softplus scores: for branch logit $g_{n,c}$, $s_{n,c} = (\mathrm{softplus}(g_{n,c}) + 10^{-6})^{1/T_d}$ and $p_{n,c} = s_{n,c} / \sum_{c'} s_{n,c'}$. At evaluation, $T_d = 1$ at the root and 1.5 below it. Leaf logits are softmax-normalized after masking illegal actions; they are not constrained to be exactly pure. The pure-action case in Figure 2 is illustrative, and the separate River deployments use pure leaves.

The cross-model prompts in Table 3 apply the action-legality mask when recomputing latent branch probabilities; the model forward pass applies it to leaf actions. The displayed leaf contributions can therefore differ from those of the model. The matched-input tests in Table 5 use forward-verified displays computed from Equation 1. Results are reported separately for these input conventions.

Numerical reconstruction and input indexing. We check whether displayed computations reproduce the ten flop models on 50 hand inputs from ten public nodes (nine board groups). Public nodes and the five legal, positive-range-weight hands per node are selected by fixed identity hashes, without target policies or model-error filtering. These boards may occur in model fitting; the check measures numerical reconstruction. Using the unchanged model classes on CPU (PyTorch 2.11.0), we capture actual NAM contributions and gates, independently reconstruct branch probabilities, all 81 leaf masses, and the final action mixture. Maximum discrepancies are 3.64×10^{-7}, 3.63×10^{-8}, and 2.38×10^{-7}, respectively. Changing the legal-action mask leaves latent routing unchanged, as required by the forward model.

Four checkpoints recompute a global Top-50 set different from their saved NAM indices; all differences occur at tied cutoff scores. The forward pass selects local channels within the recomputed set, then slices them by the saved indices. Across 400 routers, 65 therefore use only two to four NAM channels in this runtime, with five entries selected by the pre-slice mask. These counts describe the unchanged checkpoints in the stated CPU runtime. Reproducible export requires the same ordered indices and tie rule in fitting, loading, inference and display. Effective support is measured after indexing, within the model's at-most-five local input bound.

B.2 Teacher Curriculum and Training Objective

The three-phase **Hierarchical Hard Distillation** curriculum fits a locally sparse probabilistic tree to solver policies through intermediate teachers.

Phase 1: The Global Teacher (Denoising). We first train a "Global-Gated ResNet" (Teacher 1). It is restricted to $K_g = 50$ summaries globally while allowing nonlinear interactions among them.

Phase 2: The Structural Teacher (Topology). We then distill Teacher 1 into a "Deep Tree" (Teacher 2), which fits the target routing topology with higher-capacity local routers before the final local sparsity restriction.

Phase 3: Hard Student Locking (Local Support). Finally, we distill Teacher 2 into the target "Hard MDT". Crucially, we switch from soft gating to "Hard Gating" using the Straight-Through Estimator (STE). We physically lock the summary set to the Top-K (e.g., 5) at each node:

$$\mathbf{m}_{hard} = \text{TopK}(\mathbf{m}_{logits}, k = 5)$$

This restricts the student's routing support to at most five summaries per node. For reduced-width NAM checkpoints, the saved global indices and local masks must also agree; Appendix B.1 reports effective support after indexing.

Objective scaling and checkpoint selection. For the communication models, let p_i, y_i, m_i, w_i denote predictions, solver policies, legal-action masks, and hand weights. The supervised objective is

$$S_i = \sum_a m_{ia} s(p_{ia} - y_{ia}), \qquad \Delta V_i = \sum_a (y_{ia} - p_{ia}) Q^*_{ia}, \tag{12}$$

$$\mathcal{L}_{\text{base}} = \langle 50 S_i + 10[\Delta V_i]_+ + 2H_i \rangle_w \,. \tag{13}$$

Here s is unit-transition Smooth-L_1, H_i a target-probability hinge, and $\langle\cdot\rangle_w$ a within-batch weighted mean. Evaluation L_1 uses absolute action errors, reported as summed or action-averaged distances in each table. Equation 13 specifies the communication-model objective; complete training objectives are unavailable for some corpus-scale baselines. For a batch $\mathcal{B}$, $\langle z_i \rangle_w = \sum_{i\in\mathcal{B}} w_i z_i / (\sum_{i\in\mathcal{B}} w_i + 10^{-9})$. The unit-transition loss is $s(d) = d^2/2$ for $|d| < 1$ and $|d| - 1/2$ otherwise. With b_i and c_i the largest and second-largest target-probability slots,

$$H_i = \mathbf{1}[y_{ib_i} > 0.4]\,[p_{ic_i} + 0.1 - p_{ib_i}]_+. \tag{14}$$

This encourages a probability margin only for sufficiently concentrated targets; it does not rank actions by EV. The base uses summed masked Smooth-L_1, a positive-part fixed-continuation EV gap, and this hinge, with coefficients 50, 10, and 2. None is the reported validation L_1. For B complete validation batches $\mathcal{B}_b$, the checkpoint statistic is

$$e_i = \frac{\sum_a m_{ia} |p_{ia} - y_{ia}|}{\max(1, \sum_a m_{ia})}, \qquad V = \frac{1}{B} \sum_{b=1}^{B} \frac{\sum_{i\in\mathcal{B}_b} w_i e_i}{\sum_{i\in\mathcal{B}_b} w_i + 10^{-9}}. \tag{15}$$

Equal weighting of batch means need not equal a globally hand-weighted mean. The EV penalty is not the checkpoint-selection metric. The separate River experiment selects policies using exact same-tree strategic evaluation (Appendix J.2).

Distillation and auxiliary penalties. The first teacher minimizes $\mathcal{L}_{\text{base}} + R_g$. With zero-indexed epoch e, soft/hard predictions p^s, p^h, global gate values g_j, and stop-gradient sg,

$$R_g = \min(e, 50)\ \text{MSE}(p^h, \text{sg}(p^s)) + 0.05\,[\textstyle\sum_j g_j - 50]_+. \tag{16}$$

MSE averages over batch and action slots. The tree teacher and student instead use

$$\mathcal{L}_{\text{fit}} = 0.3\mathcal{L}_{\text{base}} + 2.8K_{\mathcal{B}} + r_e(R_h + R_p), \qquad K_{\mathcal{B}} = \frac{1}{|\mathcal{B}|}\sum_{i,a} t_{ia} \log \frac{t_{ia}}{p_{ia} + 10^{-9}}. \tag{17}$$

Here t is the preceding teacher's detached probability distribution, with $0 \log 0 = 0$ by continuity. Unlike the supervised base, $K_{\mathcal{B}}$ has no hand weights. The implementation's argument named `temp=2` multiplies this probability KL by four; it does not temperature-transform logits or distributions. Combined with mixing coefficient 0.7, this yields 2.8, not a softened-target objective.

For layer-mask vectors u_ℓ, leaf reaches ρ_{il}, and legally masked leaf probabilities q_{il},

$$R_h = 0.1\ \text{mean}_{\ell<k} \cos^2(u_\ell, u_k), \qquad R_p = \frac{0.5}{|\mathcal{B}|}\sum_{i,l} \rho_{il}[0.95 - \max_a q_{ila}]_+. \tag{18}$$

The hierarchy term is zero with fewer than two layers. The ramp is $r_e = \min(1, \max(0, (e-30)/d))$, with $d = 30$ for the tree teacher and $d = 50$ for the student; their soft-to-hard gate interpolation uses the same schedule. These settings describe the available communication-model implementation. Exact run-specific configurations are not fully recorded; the separate River objective is specified in Appendix J.2.

Training partitions for the 100-target study. The reference-first communication study uses ten scenario/position-specific flop models, each trained with seed 999, depth four, ternary branching, 156 input summaries, three output slots, global Top-50 selection, local Top-5 selection, and batch size 4,096. Table 8 gives partition sizes and minimum validation errors. The loader excludes all hands linked to nodes marked in preprocessing for parent-action frequency below 0.01 or either player's summed hand-range weights below 5. It shuffles the retained hand indices and reserves approximately 1% for validation, with a minimum of 10,000 for sufficiently large subsets. It does not group by public board or history, so communication situations may overlap fitting.

Table 8: **Communication-model partitions and selection statistics.** Training and allocated validation counts, and four-decimal student minima V, come from logs. Used validation counts are derived from the retained loader, not rerun measurements. The ten heterogeneous flop subsets are not seed replications or the full multi-street corpus.

Scenario and position	Training	Val. allocated	Val. used	Best V
BTN vs BB, 3-bet, IP	694,712	10,000	8,192	0.1017
BTN vs BB, 3-bet, OOP	1,668,420	16,853	16,384	0.1091
BTN vs BB, single-raised, IP	1,184,238	11,963	8,192	0.1249
BTN vs BB, single-raised, OOP	2,007,763	20,281	16,384	0.0551
BTN vs SB, 3-bet, IP	697,960	10,000	8,192	0.1024
BTN vs SB, 3-bet, OOP	1,096,513	11,076	8,192	0.1061
SB vs BB, 3-bet, IP	1,014,509	10,248	8,192	0.1365
SB vs BB, 3-bet, OOP	752,443	10,000	8,192	0.0646
SB vs BB, single-raised, IP	2,179,277	22,013	20,480	0.1114
SB vs BB, single-raised, OOP	2,708,003	27,354	24,576	0.0916

The validation loader does not shuffle and drops each incomplete final batch. Its batch size implies 126,976 records used per pass out of 149,788 allocated, with 22,812 omitted across the ten models. These counts are reconstructed from the available implementation and training logs; exact per-run validation indices are unavailable. Each minimum selects a checkpoint on the same validation partition and is therefore a selection statistic, not an independent test error. The corpus-scale comparison has separate evaluation and dispersion conventions.

Training settings. The available communication-model implementation uses AdamW, with epoch caps 150, 150, and 250; the first teacher can stop early on its validation criterion. Teacher 1 uses learning rate 3×10^{-4} for gate and backbone. Teacher 2 uses 10^{-4} for its global gate, 3×10^{-4} for local gates and routers, and 10^{-3} for leaves. The student fixes global selection, uses 3×10^{-3} for routers and 10^{-3} for local gates and leaves, and uses cosine scheduling with a 10^{-5} floor. Complete per-run baseline settings and the replication units behind the corpus tables' $\pm$ entries are unavailable; those entries cannot be interpreted as confidence intervals from these ten models.

C Policy Fidelity and Local Input Support

C.1 Grouped Holdout with Matched Local Support

Question and sampling unit. This controlled experiment asks whether routing information transfers to public boards excluded from fitting, and measures the effect of local input support while holding student architecture and optimization fixed. We use the five NLH scenarios and two acting seats, with a separate model for each of the ten subsets. Flop boards equivalent under any permutation of suits belong to one group across all scenarios, seats and betting histories. A fixed group partition is shared by all models. The selected training, validation and test sets contain 256, 64 and 64 distinct groups, respectively, with no group overlap.

Within each subset and partition, a seeded identity hash chooses one eligible node per group and then the required number of groups. Eligibility requires at least sixteen positive-reach hands and two legal output slots, not a mixing, policy-distance or model-agreement condition. Sixteen distinct hands per node are sampled uniformly without replacement. Each subset therefore has 4,096 training, 1,024 validation and 1,024 test rows. The test set contains 640 nodes but only 64 independent board groups. Normalization and feature selection use training rows only. The 156-column input representation and its missingness masks include the oracle-derived summaries used elsewhere in the paper. This tests new-context prediction with those summaries available.

Matched fitting and checkpoint selection. All teachers are randomly initialized. For each subset and seed, a common globally gated ResNet and deep-router tree are fitted before the student comparison. The teacher epoch caps are 150 and 150; the first teacher uses the same early validation threshold. Students use the same depth-four ternary topology, forty additive routers of width sixteen, fifty globally selected inputs and eighty-one leaves. Only the local support budget changes: $K = 5, 10, 50$. All three students share identical initial parameters, the selected teacher, data, masks, batch order, loss coefficients, optimizer and 250-epoch schedule. The local mask is soft through epoch 29, ramps to hard by epoch 80, and remains hard thereafter. The single 4,096-row batch gives equal update counts. Training uses the reach-weighted distillation objective in Appendix B.2; each model selects its minimum uniform validation summed-L_1 checkpoint under hard-mask inference. No test score selects an epoch, seed or support budget.

A stable feature-index tie rule is used for global and local Top-K selection and agrees with the saved NAM input order. All three student variants contain 172,951 stored parameters; their global teacher and tree teacher contain 8,571,351 and 4,318,111, respectively. Thus the student comparison varies local input support without changing parameter capacity.

Training seed 20260922 is the primary panel; 20260923 and 20260924 are fixed replications on the same data. Per-seed results and their equal-seed aggregate are reported separately. This controlled subset of the corpus isolates local support under matched fitting. Checkpoint selection and support verification use only training and validation data.

Routing and input-distance controls. The first four identity-sampled hands at each test node supply fixed reference policies; the remaining twelve are queries, giving 7,680 queries per model. Numerical prediction is a convex combination of the same four reference labels. The primary kernel is the histogram intersection in Equation 4. A separate dot-product kernel, $\sum_l \rho_l(\mathbf{x}_h)\rho_l(\mathbf{x}_{h_j})$, is retained as a sensitivity condition. Their weights are normalized across references, with uniform fallback at zero total weight. Each has a fixed control that cyclically shifts the four weights by one place while leaving reference labels unchanged.

Input-distance controls (feature distances) use summed absolute differences over jointly present columns, either in original units or divided by training-only standard deviations. The latter excludes training-constant columns. An incomparable reference has zero weight; if all are incomparable, use uniform weights. Exact comparable matches share the weight equally; otherwise weight by $(d + 10^{-6})^{-1}$. No query policy, model action output or leaf policy enters these weights. All queries remain in every condition.

Display and uncertainty. For each query and student, fixed retained-mass thresholds of 50, 75, 90, 95, 99 and 100% define descending prefixes, breaking ties by leaf index. We report error to the complete MDT and to the solver, retained leaves, ancestor routers, and the union of their declared input identities. This conservative static footprint counts an input even if missing in a particular query. Complete mixture replay must agree with the forward output within 2×10^{-6}; every partial display must satisfy the retained-mass error bound at that tolerance. These are computational footprint measures, not human timings.

Same-local-budget post-hoc control. An inference-only control loads the paired Top-50 student's unchanged selected checkpoint and applies a hard Top-5 mask. This asks whether simply imposing the local budget after fitting achieves the fidelity of training with that budget. No parameters, normalization statistics, checkpoints or thresholds are adapted. Both models have five inputs per router; their whole-display input unions can differ and are measured separately. This is a fixed masking baseline, not an optimized pruning procedure. Its checkpoint was selected using Top-50 validation error.

Errors are averaged within a node and then uniformly across nodes, with a reach-weighted sensitivity result. Paired bootstrap samples resample the 64 canonical board groups, retaining their repeated appearances across subsets. Seven contrasts use 20,000 draws and Bonferroni-adjusted 99.2857% intervals: Top-5 minus Top-10 and Top-50 policy error, standardized-feature minus each kernel's interpolation error, and shuffled minus correct assignment for each kernel, plus post-hoc-mask minus trained Top-5 policy error. Both kernels and the post-hoc control were specified before test scoring; the seven-contrast correction covers the joint results. Bootstrap uncertainty is conditional on these fitted seeds and this split; per-seed replication is reported separately, rather than treating seeds as additional independent boards.

Held-out fidelity and fixed-budget prediction. Table 9 reports every fitted model and seed. All-hand scores include sixteen sampled hands at each node. The numerical reference experiment instead uses twelve queries, keeping the other four hands as references; its action-head baseline is scored on those same queries (Table 10). The two populations are not pooled. The three-seed mean routing gain over standardized features is 0.05858 summed L_1, or 9.76%. Both kernels preserve useful same-context information on boards excluded from fitting, supporting routing as an informative intermediate representation. The complete action head remains more accurate than reference interpolation.

The matched student comparison isolates the practical effect of a native local budget while keeping parameter capacity fixed. Top-5 has higher error than Top-10 and Top-50, but substantially lower error than simply applying a Top-5 mask to the selected Top-50 checkpoint. The latter comparison includes native checkpoint selection as well as training with the intended budget; it does not isolate optimization from checkpoint selection or compare against the best possible pruning algorithm. All seven primary-seed intervals exclude zero in their reported direction, and the two replications agree (Table 11).

Table 11: **Seven jointly adjusted contrasts.** Paired board-cluster bootstrap, 20,000 draws. Each interval has 99.2857% nominal coverage (seven-contrast Bonferroni correction). Seed panels are replications on one partition; Mean averages seeds within each board before resampling. Positive differences favor the second named method.

Seed	Contrast	Difference	Adjusted interval
22	Top-5 minus Top-10	0.0343	$[0.0287, 0.0400]$
22	Top-5 minus Top-50	0.0656	$[0.0580, 0.0731]$
22	Std. features minus dot	0.0681	$[0.0531, 0.0840]$
22	Permuted minus dot	0.1941	$[0.1659, 0.2259]$

Seed	Contrast	Difference	Adjusted interval
22	Std. features minus intersection	0.0584	[0.0456, 0.0715]
22	Permuted minus intersection	0.1751	[0.1493, 0.2038]
22	Post-hoc mask minus trained Top-5	0.2956	[0.2757, 0.3146]
23	Top-5 minus Top-10	0.0369	[0.0312, 0.0428]
23	Top-5 minus Top-50	0.0740	[0.0662, 0.0814]
23	Std. features minus dot	0.0695	[0.0548, 0.0843]
23	Permuted minus dot	0.1905	[0.1621, 0.2210]
23	Std. features minus intersection	0.0584	[0.0460, 0.0710]
23	Permuted minus intersection	0.1724	[0.1465, 0.1994]
23	Post-hoc mask minus trained Top-5	0.3328	[0.3082, 0.3559]
24	Top-5 minus Top-10	0.0375	[0.0307, 0.0446]
24	Top-5 minus Top-50	0.0711	[0.0616, 0.0801]
24	Std. features minus dot	0.0671	[0.0518, 0.0820]
24	Permuted minus dot	0.1878	[0.1579, 0.2202]
24	Std. features minus intersection	0.0589	[0.0459, 0.0719]
24	Permuted minus intersection	0.1742	[0.1475, 0.2033]
24	Post-hoc mask minus trained Top-5	0.3149	[0.2939, 0.3354]
Mean	Top-5 minus Top-10	0.0362	[0.0320, 0.0406]
Mean	Top-5 minus Top-50	0.0702	[0.0635, 0.0769]
Mean	Std. features minus dot	0.0682	[0.0538, 0.0827]
Mean	Permuted minus dot	0.1908	[0.1626, 0.2218]
Mean	Std. features minus intersection	0.0586	[0.0463, 0.0710]
Mean	Permuted minus intersection	0.1739	[0.1481, 0.2019]
Mean	Post-hoc mask minus trained Top-5	0.3144	[0.2942, 0.3335]

Display trade-off and verification. Figure 3 and Table 12 show every fixed retained-mass budget. At 95%, Top-5 uses 39.03 distinct input identities, compared with 46.33 for Top-10 and 50 for Top-50; its average discrepancy to its own complete policy is 0.02576. The corresponding router and leaf counts are larger than Top-50's. Thus a local input constraint changes the structure of the computation, rather than shrinking every measure of display size. Truncating a mixture can sometimes reduce its solver error while increasing its discrepancy from the complete model. No retained-mass threshold is selected by test solver error.

Group disjointness is checked before fitting. Numerical validation checks parameter counts, legal policies and forward reconstruction. An independent scalar implementation rescored 69,120 reference-query records and 414,720 partial displays, plus 30,720 post-hoc-mask predictions and 138,240 associated displays. Every display met the retained-mass bound and the full-mixture reconstruction tolerance. The supplementary implementation provides reference weighting and display calculations with synthetic tests. Training used one RTX 5090 sequentially for 1.003 GPU hours; numerical scoring used the CPU.

C.2 Corpus-Scale Architecture Comparison

Table 14 reports action-distribution L_1 and oracle-conditioned EV gap for solver-policy distillation. Dense high-capacity models obtain the lowest numerical loss. The raw-PBS baselines use a 1482-dimensional input and 512-wide hidden layers, whereas the summary-input baselines use a 156-dimensional input and 256-wide hidden layers. The soft-sparse tree keeps dense routers over these summaries and uses regularization rather than hard selection; Teacher 2 uses higher-capacity two-layer MLP routers; the final hard student applies local Top-5 masks. The tree variants improve on the summary-input flat MLP, but the final hard student's higher error does not isolate sparsity as its cause: these families differ in capacity and training. Hard selection bounds computational input support; it does not ensure that a separate renderer names the same inputs or reproduces their effects (Appendix B.1).

C.3 Communication-Model Fidelity on Sampled Nodes

Population and selection. This evaluation measures the fidelity of the ten seed-999 scenario/seat models used for communication. The sampling pool is the available preprocessed flop data for five

Table 9: **Every training seed: full policy fidelity.** Summed action L_1 over 10,240 test hands per seed. Seed suffixes denote 20260922–20260924; seed 22 is primary. SD is across seeds, not a confidence interval.

Model	Seed 22	Seed 23	Seed 24	Mean	SD
Global-gated ResNet	0.2802	0.2790	0.2821	0.2804	0.0016
Tree teacher	0.3993	0.3998	0.3931	0.3974	0.0038
Top-5	0.4247	0.4277	0.4230	0.4251	0.0024
Top-10	0.3904	0.3908	0.3855	0.3889	0.0030
Top-50	0.3590	0.3537	0.3519	0.3549	0.0037
Top-50→5	0.7202	0.7605	0.7379	0.7395	0.0202

Table 10: **Fixed-reference prediction on grouped holdout.** Summed action L_1, 7,680 queries per seed. All interpolators use the same four labels. The last row uses the full trained action head and is not an information-matched reference method.

Method	Seed 22	Seed 23	Seed 24	Mean
Raw features	0.6008	0.6008	0.6008	0.6008
Standardized features	0.6002	0.6002	0.6002	0.6002
Histogram intersection	0.5418	0.5419	0.5413	0.5416
Permuted intersection	0.7169	0.7142	0.7155	0.7155
Dot product	0.5321	0.5307	0.5332	0.5320
Permuted dot product	0.7262	0.7211	0.7209	0.7227
Full MDT action head	0.4255	0.4280	0.4239	0.4258

preflop scenarios and both seats. A node is eligible when at least two abstract actions are legal and at least 16 hand combinations have positive acting-player reach. Before reading target policies, nodes are ranked by SHA-256 of a fixed seed (202609213), subset, board, history, player and node identity. We select ten nodes per subset, excluding the 98 distinct exact boards of the reference-first 100-target communication experiment and requiring distinct boards across all 100 nodes. Within each node, 16 hands are sampled uniformly without replacement from positive-reach combinations. Selection uses no policy mixing, solver distance, model agreement, reference diversity or prediction error.

Every node contributes 16 equally weighted hands; the mean is not reach-weighted. These boards differ from the reference-first communication cases but may overlap MDT fitting. The population is limited to the available preprocessed data. The architecture comparison in Appendix C.2 uses a separate corpus-scale evaluation.

Models and numerical checks. Each subset uses its retained globally gated ResNet teacher, higher-capacity tree teacher and hard sparse student. The first teacher uses 50 global inputs; the tree teacher uses a 50-input global gate and up to eight local inputs. The student preserves the checkpoint's ordered 50-input index tensor and at most five effective inputs per router. Global-cutoff ties can leave fewer effective local inputs (Appendix B.1). The respective stored parameter counts are 8,571,351, 4,318,111 and 172,951. The student has 40 ternary routers and 81 soft paths; parameter count includes its gates and leaf policies, not only a printed route.

All models run in evaluation mode on CPU. An independent forward computation agrees with all 1,600 student predictions within 8.35×10^{-7}. Reconstructed additive router scores, normalized probabilities, path products and leaf mixtures meet a 2×10^{-6} tolerance. Model parameters remain fixed throughout this evaluation.

Metrics and uncertainty. Summed L_1 adds the absolute errors in check/call, fold and bet/raise; slot L_1 divides this sum by three, including illegal slots fixed to zero. Table 16 uses 50,000 bootstrap samples of the 100 board-level means, seed 202609213. The intervals quantify sampled-board uncertainty for these fixed checkpoints, not training-seed variation. The teacher/student comparison includes architecture, capacity and training differences; it does not isolate the effect of sparsity alone. Table 1 tests local support under matched fitting on held-out board groups.

Table 12: **All fixed display budgets.** Equal-seed means on the query-only population. Model L_1 measures discrepancy to the same model before truncation; Inputs counts distinct declared input identities. No budget is chosen using these test scores.

Model	Mass (%)	Solver L_1	Model L_1	Inputs	Routers	Leaves
Top-5	50	0.4033	0.2299	25.33	7.62	3.09
Top-5	75	0.3946	0.1126	31.31	11.55	6.38
Top-5	90	0.4068	0.0480	36.37	16.17	11.75
Top-5	95	0.4146	0.0258	39.03	19.40	16.60
Top-5	99	0.4227	0.0067	42.39	25.38	29.18
Top-5	100	0.4258	0.0000	46.13	40.00	81.00
Top-10	50	0.3770	0.2335	36.32	6.69	2.43
Top-10	75	0.3612	0.1229	41.18	9.87	4.72
Top-10	90	0.3706	0.0524	44.65	13.59	8.34
Top-10	95	0.3784	0.0261	46.33	16.29	11.64
Top-10	99	0.3870	0.0057	48.28	21.90	21.12
Top-10	100	0.3897	0.0000	49.67	40.00	81.00
Top-50	50	0.3533	0.2760	50.00	5.30	1.65
Top-50	75	0.3232	0.1340	50.00	7.28	2.79
Top-50	90	0.3331	0.0546	50.00	9.57	4.35
Top-50	95	0.3425	0.0271	50.00	11.25	5.63
Top-50	99	0.3530	0.0050	50.00	14.95	9.06
Top-50	100	0.3556	0.0000	50.00	40.00	81.00
Top-50→5	50	0.6664	0.2198	35.17	14.51	8.99
Top-50→5	75	0.6972	0.1108	41.80	22.87	21.64
Top-50→5	90	0.7191	0.0516	44.87	29.85	37.97
Top-50→5	95	0.7280	0.0285	45.92	33.24	48.34
Top-50→5	99	0.7362	0.0064	46.88	37.33	64.86
Top-50→5	100	0.7385	0.0000	47.30	40.00	81.00

Table 13: **Reach-weighted sensitivity.** Within each node, weight the sixteen sampled hands by normalized reach; then average nodes uniformly. This is conditional on the sampled hands, not a full-range census.

Model	Seed 22	Seed 23	Seed 24
Global-gated ResNet	0.2638	0.2611	0.2631
Tree teacher	0.3730	0.3746	0.3671
Top-5	0.3992	0.4007	0.3977
Top-10	0.3657	0.3642	0.3633
Top-50	0.3334	0.3287	0.3276

Reproducibility. The evaluation records identify sampled hands and all 30 teacher/student checkpoints, with per-hand predictions and independent routing reconstruction. Access to solver-derived inputs remains subject to the data license. The intervals here use explicitly defined board resampling; they are separate from the unspecified dispersion convention of the corpus-scale architecture comparison.

D Mixture Displays and Input Definitions

D.1 Single Paths and Retained-Mass Mixtures

A locally greedy route need not reach the globally most probable leaf, and either leaf is only one component of Equation 1. We measure how much policy information a single path omits on all 100 reference-first targets. All 81 legal-action-masked leaf prototypes and their reaches reconstruct the full MDT output within 2×10^{-6}. This analysis measures fidelity to the model's computation, separately from the LLM prediction tests.

Table 14: **Distillation fidelity on the NLH solver-labeled evaluation set.** Lower L_1 and oracle-conditioned EV gap are better. The oracle row is zero by construction; the gap is not full-game exploitability. The recorded means and $\pm$ entries are reported, but the dispersion definition and replication units could not be recovered from the retained records. No confidence interval or significance claim is inferred from those entries.

Input / group	Model	L_1	EV gap (% pot)	Architecture / params	Visibility
-	*Oracle*	-	*0.00%*	-	*N/A*
Raw PBS	2-Layer MLP	0.068 ± 0.002	$0.55 \pm 0.04\%$	512 hidden; ∼1.02M	Black-box policy
	8-Layer ResNet	0.036 ± 0.001	$0.18 \pm 0.02\%$	512 hidden; ∼2.60M	Black-box policy
Strategic summaries	1-Layer MLP	0.098 ± 0.010	$0.72 \pm 0.05\%$	256 hidden; ∼41K	Dense summary model
	2-Layer MLP	0.053 ± 0.005	$0.37 \pm 0.03\%$	256 hidden; ∼107K	Dense summary model
	4-Layer MLP	0.045 ± 0.005	$0.26 \pm 0.03\%$	256 hidden; ∼238K	Dense summary model
	8-Layer ResNet	0.021 ± 0.003	$0.05 \pm 0.01\%$	256 hidden; ∼502K	Dense summary model
Tree students	Dense-router tree	0.031 ± 0.004	$0.12 \pm 0.03\%$	∼19K	Tree, dense routing
(Ablations)	Soft-sparse tree	0.057 ± 0.005	$0.21 \pm 0.02\%$	Sparsity reg.; ∼19K	Tree, regularized routing
Hard curriculum	Teacher 1 (global sparse)	0.028 ± 0.002	$0.07 \pm 0.01\%$	8-layer ResNet; ∼502K	Global summary gate
	Teacher 2 (local sparse)	0.044 ± 0.005	$0.23 \pm 0.03\%$	2-Layer MLP routers; ∼1.64M	Local tree routing
	Final hard MDT	0.087 ± 0.007	$0.44 \pm 0.04\%$	top-k/node mask; ∼0.7K	Explicit Top-5

Table 15: Summed action L_1 by scenario/seat subset; 160 hands and ten distinct boards per row. The same hands are used for all three models.

Subset	Global teacher	Tree teacher	Hard MDT
BTN vs BB 3-bet, IP	.0466	.4638	.3463
BTN vs BB 3-bet, OOP	.1260	.2693	.3344
BTN vs BB SRP, IP	.3013	.2772	.3252
BTN vs BB SRP, OOP	.0661	.1560	.1654
BTN vs SB 3-bet, IP	.0429	.4691	.3143
BTN vs SB 3-bet, OOP	.1031	.1956	.2488
SB vs BB 3-bet, IP	.1371	.2187	.2781
SB vs BB 3-bet, OOP	.0613	.1217	.1772
SB vs BB SRP, IP	.0466	.1664	.2232
SB vs BB SRP, OOP	.0783	.1763	.2635

A retained-mass guarantee. For nonempty leaves S with total reach $m = \sum_{l \in S} w_l > 0$, display the normalized partial mixture $p_S = \sum_{l \in S} w_l q_l / m$. Writing $p = m p_S + (1 - m) p_{\bar{S}}$ gives

$$\|p - p_S\|_1 = (1 - m)\|p_{\bar{S}} - p_S\|_1 \leq 2(1 - m). \tag{19}$$

For $m = 1$ the discrepancy is zero. The inequality follows from the maximum distance between probability vectors; it is an elementary worst-case bound, not a novel learning theorem. It can be tight for disjoint-support prototypes, while identical prototypes give zero error even at small retained mass. We sort by descending reach with leaf-ID tie-breaking, report fixed prefix sizes, and separately find the smallest prefix reaching 95% mass. Greedy local argmax and highest global reach are kept distinct. The local greedy leaf differs from the globally highest-reach leaf in 51/100 targets and

Table 16: **Communication-checkpoint fidelity.** Same 1,600 sampled hands on 100 separately sampled exact boards, without certified exclusion from fitting. Intervals are 95% board-bootstrap intervals for summed L_1.

Model	Summed L_1	Slot L_1	95% interval
Global-gated ResNet	0.1009	0.0336	$[0.0799, 0.1255]$
Tree teacher	0.2514	0.0838	$[0.2231, 0.2808]$
Hard Top-5 MDT	0.2676	0.0892	$[0.2451, 0.2904]$

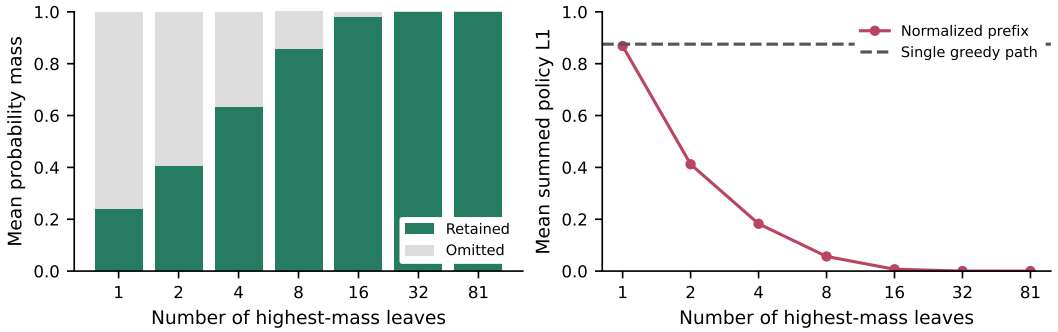

Figure 4: **Information omitted by a path display.** All 100 target averages. Left: mass retained by highest-reach leaves. Right: summed discrepancy from the complete MDT mixture; the dashed line is a locally greedy single-leaf prototype. This is not an LLM result.

from the complete mixture's dominant action in 32/100. Its mean solver error is 0.928329 versus 0.204340 for the complete MDT. A 95% mass prefix uses 11.97 leaves on average, with mean MDT discrepancy 0.017368.

Validation. All prefixes satisfy the bound; the complete prefix reconstructs the policy. Independent tests cover random mixtures, tight and zero-error cases, and the distinction between local and global path maxima. Small floating-point reach error is normalized for this bound. The complete target MDT policy is used here to measure reconstruction, but remains hidden from all reported LLM conditions.

D.2 Displays with a Prescribed Error Budget

The bound in Appendix D.1 allows a display to meet a prescribed approximation error. We sort leaf reaches, break ties by leaf index, and include the shortest prefix with mass at least $1-\epsilon/2$. The display contains leaf indices, reaches, prototypes, legal actions, retained and omitted mass, the normalized partial mixture, and its bound $2(1-m)$. Independent recomputation checks normalization, legal probability mass and consistency with the complete mixture. Source rounding within 2×10^{-6} is normalized explicitly.

Validation and reproducibility. Five test groups cover synthetic mixtures, exact thresholds, ties, zero tolerance, invalid inputs and altered fields. A separate scalar implementation checks all 100 saved target certificates without importing the display generator. At $\epsilon = 0.1$, all 100 satisfy the bound. The mean display has 11.97 leaves; mean summed L1 to the complete MDT is 0.017368, maximum observed error 0.056148, and maximum certified bound 0.099955. These are computational checks of the same models, not new independent test decisions.

Table 17: **Full-population display fidelity.** Prefix policies are normalized; all errors are summed action L_1.

Display	Mean retained mass	Mean L_1 to MDT	Mean upper bound
Local greedy leaf	0.207462	0.875330	1.585077
Top 1 leaves	0.241817	0.868166	1.516367
Top 2 leaves	0.407298	0.412016	1.185403
Top 4 leaves	0.634683	0.182437	0.730633
Top 8 leaves	0.856609	0.056489	0.286782
Top 16 leaves	0.981158	0.007050	0.037684
Top 32 leaves	0.999744	0.000112	0.000513
Top 81 leaves	1.000000	0.000000	0.000000

Table 18: **All scenario/seat groups.** Ten targets per row; greedy means the local-argmax leaf, and p the full mixture.

Group	Greedy mass	Greedy–p L_1	p–solver L_1	95% leaves
BTN_vs_BB_3B_IP	0.1898	0.9643	0.3171	13.3
BTN_vs_BB_3B_OOP	0.2584	0.7940	0.2292	10.0
BTN_vs_BB_SRP_IP	0.1787	0.9957	0.1446	12.4
BTN_vs_BB_SRP_OOP	0.2948	0.6232	0.2692	8.9
BTN_vs_SB_3B_IP	0.2507	0.7219	0.2214	10.6
BTN_vs_SB_3B_OOP	0.2134	0.8466	0.1821	10.7
SB_vs_BB_3B_IP	0.1641	1.0053	0.1961	15.3
SB_vs_BB_3B_OOP	0.1730	0.9132	0.1426	11.0
SB_vs_BB_SRP_IP	0.1354	0.9385	0.1848	14.9
SB_vs_BB_SRP_OOP	0.2163	0.9507	0.1563	12.6

Table 19: **Every subset under the fixed 0.1 error budget.** Ten targets per row; errors are to MDT, not the solver.

Subset	Mean leaves	Mean summed L1 to MDT
BTN_vs_BB_3B_IP	13.30	0.013797
BTN_vs_BB_3B_OOP	10.00	0.017703
BTN_vs_BB_SRP_IP	12.40	0.020102
BTN_vs_BB_SRP_OOP	8.90	0.018431
BTN_vs_SB_3B_IP	10.60	0.012993
BTN_vs_SB_3B_OOP	10.70	0.020585
SB_vs_BB_3B_IP	15.30	0.011343
SB_vs_BB_3B_OOP	11.00	0.021250
SB_vs_BB_SRP_IP	14.90	0.021301
SB_vs_BB_SRP_OOP	12.60	0.016180

The supplementary implementation and synthetic tests require no licensed data or weights; reproducing the reported aggregates requires licensed inputs. These target displays are used only to measure approximation of MDT output and are not supplied to the LLM.

D.3 Whole-Display Input and Path Budgets

Within-model comparison. To quantify the reconstruction budget in Appendix D.1, we measure reconstruction error and the number of distinct displayed inputs on all 300 target-first queries from Appendix G.7. The ten models, inputs and masks are fixed. We compare greedy local routing, the globally largest-reach leaf, and shortest descending prefixes retaining 50%, 75%, 90%, 95%, 99% or 100% mass. All thresholds are fixed before scoring; only the displayed sub-mixture changes.

For each retained set, normalize its path masses and mix its legal leaf policies. The outcome is summed L_1 distance to the complete MDT policy, not solver error or LLM prediction accuracy. Count retained leaves, all visited routers and distinct effective input identities. Five inputs per router

is a local computational constraint; repeated router visits and overlapping input sets determine the whole-display footprint.

Table 20: Display frontier on the same 300 queries. Footprints and discrepancies are means; the final column gives the largest discrepancy. All errors compare against the complete MDT, not the solver.

Display	Leaves	Routers	Inputs	Mean L_1	Max L_1
Greedy route	1.00	4.00	16.11	.421640	1.495485
Largest-mass path	1.00	4.00	16.10	.396967	1.287431
50% mass	2.54	6.95	22.30	.201241	.925780
75% mass	4.67	9.74	26.12	.108986	.481416
90% mass	7.29	12.30	28.85	.047583	.195585
95% mass	9.11	13.77	29.95	.024627	.078966
99% mass	13.26	16.75	31.54	.005444	.018906
Full tree	81.00	40.00	37.10	.000000	.000000

Retained-mass bound. Writing the complete distribution as $p = mp_S + (1 - m)p_{\bar{S}}$ gives the elementary bound $\|p - p_S\|_1 \leq 2(1 - m)$ from Appendix D.1. Every one of the 2,400 trace records satisfies its own retained-mass bound, with 2×10^{-6} numerical tolerance. At 95% mass, the worst-case guarantee is 0.1; the largest observed error is 0.078966. The average retained mass is 0.962712. The display uses at most twenty leaves, twenty-eight routers and forty-four distinct inputs in this sample. Across checkpoints, the full tree averages 37.10 distinct summaries, with a maximum of forty-nine.

The supplementary implementation checks the normalized partial policy, omitted-mass bound and number of distinct inputs. Tests cover a disjoint-support example attaining the bound, complete reconstruction and invalid probabilities. Independent forward reconstruction verifies the path and leaf probabilities, while scalar calculations reproduce the reported aggregates and every mass threshold.

Interpretation. This frontier quantifies the trade-off between approximation accuracy and display size. The 95% display uses fewer leaves and about thirty distinct summaries. These counts cover one target's trace; the full five-hand prompt includes additional reference evidence. The calculation uses the complete leaf distributions and measures numerical reconstruction. LLM use of the displayed evidence is evaluated separately in Appendix G.1.

D.4 Common Input Definitions and Withholding

The withholding rules in this subsection apply to the full-reference-path comparison in Appendix G.2; other comparisons specify their own input sets. All 40 prompts match their source values and paired construction. Table 21 enumerates all 37 fields using their zero-based input column IDs. Definitions distinguish source-derived quantities from independently reconstructed poker quantities, providing the semantic distinctions used in forward-verified displays.

Withheld fields and missing values. The 19-field rule comprises nine policy/value-dependent fields, five fields with unresolved formula or semantic interpretation, two opaque provider scores, two synthetic-equity fields without the required upstream source binding, and one padding field. Source masks additionally mark 24 target cells missing: columns 135 and 136 in 11 cases each, and 132 and 146 in one case each. These reasons overlap the withholding rule; they add no nulls among the 18 emitted fields. In all 20 cases, five withheld node summaries (075, 079, 083, 085, 090) are nevertheless visible unchanged through every reference row, as is the zero padding value. Both prompt conditions therefore retain the same oracle evidence, including some target-related information.

Input consistency checks. On all 100 selected target/reference rows, column 152 agrees with a direct card-identity calculation to within 2.56×10^{-8}. Its historical name contains *kicker strength*, but the quantity is only lower hole-card rank divided by 14, not best-five-hand kicker strength. Columns 082, 124 and 137 satisfy their internal difference/average identities within 1.50×10^{-8}; column 143 agrees exactly with the same-source selected hand weight. Those four checks verify internal identities

Table 21: **The 37 shared field definitions for the full-reference-path test.** All 18 emitted target fields have values in all 20 cases. The other 19 are null in every target row; reference rows retain source-present values in both conditions. EQ denotes stored equity. Named provider categories are retained as provider labels, not recertified best-five-card classes.

Column IDs	Stored meanings	Target treatment
001, 005	Second history-action code; OOP stack normalized by matched pot	Emitted
017, 053	OOP SecondPair range mass; IP Set range mass	Emitted
082	OOP minus IP range equity	Emitted
091, 117	AceHigh indicator; NoDraw indicator	Emitted
122, 123	Equity of opponent/own range hands sharing one private card	Emitted
124–126	Own-range card frequencies: average, first card, second card	Emitted
128	Current hand equity	Emitted
137, 139, 140	Hand EQ minus nut threshold; EQ minus range mean; blocker equity minus its range mean	Emitted
143, 152	Normalized hand range weight; lower private-card rank divided by 14	Emitted
079, 083, 085, 090	Descendant bluff efficiency; OOP value share; value-share difference; bet-range morphology	Withheld: policy/value
127, 131, 132, 145, 146	Hand EV; bet–call and bet–raise equity; two conditional response-line regrets	Withheld: policy/value
134–136, 141, 142	Signed vulnerability; history-derived MDF; EQ minus MDF; relative vulnerability; equity rank	Withheld: semantics
120, 121	Provider blocker and unblocker scores	Withheld: opaque definition
075, 138	Synthetic top-range equity difference; hand EQ minus synthetic-range mean	Withheld: unbound source
059	Unmapped padding slot, observed zero	Withheld: placeholder

in the stored quantities rather than independently validating upstream equity, ranges or normalization. The remaining 13 emitted fields are source-traced to historical arrays rather than reconstructed from raw inputs. The 18-field vocabulary is therefore reported with its source-defined semantics and the withholding rules above.

E Same-Context Evidence and Communication Protocols

E.1 Evaluation Questions and Populations

MDT is trained on solver policies; downstream LLM weights are fixed. The appendix separates three scientific questions:

Representation and fidelity. Appendices B.2, C.1 and D.1 specify fitting, board-held-out prediction, local support and numerical display error. Communication-checkpoint fidelity is measured separately in Appendix C.3.

In-context strategy communication. SCCS constructs the evaluation evidence. Package tests (Appendices F.1, F.2, F.5, F.6) differ from information-matched path tests (Appendices G.1 and G.4) and numerical reference selection/decoding (Appendices H.1 and H.6). Each retains its sampling unit, renderer, coverage and uncertainty convention.

Strategic validation and portability. River best responses evaluate complete in-domain policies (Appendix J.2); Liar's Dice tests private-type cross-fitting and a second-game communication interface (Appendix K.1).

The cross-model 100-query result uses policy-filtered targets, parsed outputs, and the input conventions documented below. Its 52.6% reduction and the forward-verified and target-first estimates

each refer to their stated populations. Prediction accuracy, numerical fidelity and strategic value are reported separately.

E.2 Reference-First Sampling and Target Filters

SCCS supplies same-context evidence for fixed-LLM evaluation of the learned MDT representation. The reference-first NLH procedure selects a reference and three shadows using model-policy differences, then filters an undisplayed target using solver labels. Algorithm 1 specifies this selection; the target-first protocol is in Appendix F.1.

Algorithm 1 Scenario-Constrained Counterfactual Sampling (SCCS): Reference-First Evidence Construction

Require: Reference hand h_0, trained MDT $\mathcal{M}$, same-context hand pool $\mathcal{D}$
Ensure: Four displayed references and one undisplayed evaluation hand, or no sample

1: Compute h_0's model policy and greedy route.
2: Use all same-node candidates, or sample 2,000 if the pool is larger.
3: Remove h_0 and candidates with model summed-policy distance below 0.05.
4: Find the first divergence from the reference's greedy route.
5: If no divergence exists, discard the candidate when its maximum-reach leaf action equals the reference policy's dominant action; otherwise assign depth D.
6: Rank by later divergence, then smaller raw-feature summed distance.
7: Greedily select three shadows, requiring model summed-policy distances at least 0.03 between selected shadows.
8: Choose an undisplayed remaining hand, preferring a leaf action different from the reference's dominant action.
9: Require the target solver maximum probability ≤ 0.90 and second probability ≥ 0.08.
10: Require nearest-reference solver summed distance in $[0.02, 0.35]$ and target model–solver summed distance ≤ 0.35.
11: Require reference maximum/mean pairwise solver summed distances $\geq 0.30/0.10$, and at least one reference with second probability ≥ 0.05.
12: Exclude repeated evaluation hands and public nodes; otherwise retain the sample.
13: Display reference features, traces, model/solver policies, and target feature evidence, withholding target solver policy and aggregate MDT mixture.
14: Ask the fixed predictor for its JSON action distribution; score it locally.

All thresholds in Algorithm 1 are *summed* policy distances; reported errors average over the three action slots. The builder records seed 20260425 and at most 5,000 reference attempts per subset. The selected evaluation distribution favors mixed, locally predictable cases, not random solver decisions. Target exclusion refers to the displayed references and does not establish exclusion from MDT training.

Sample characteristics. The reference-first 100-target evaluation contains four same-context references and a distinct evaluation hand at each of 100 public nodes. All 100 pass the nearest-reference summed-distance filter; its observed range is 0.02407–0.34407, with median action-averaged distance 0.04139. These distances define the selected population in Algorithm 1. The illustrative case in Appendix F.7 is separate from this scored set. Figure 5 shows a forward-verified input from the compact-reference test in Appendix G.3.

E.3 Selection Yield and a Verified Contrast

Selection yield characterizes the population induced by the reference-first filters. Because per-attempt event logs are unavailable, we reconstruct the candidate sequence using seed 20260425, the input lengths and subset order. This reproduces all 100 selected reference indices in order, with no repeated accepted index in its processed prefix. Thirty index, weight and mask arrays are verified against the evaluation inputs. The yield is a reconstruction from these inputs and the sampling implementation, not a recorded event count.

The stream processes 983 attempts to reach ten samples per subset; ten unused quota-stop lookahead draws are excluded. Two attempts fail reference weight 10^{-6}; none fail the two-legal-action or

retained-node tests. Of 981 initial passes, 100 are retained. The remaining 881 cannot be assigned to shadow construction, target filters, duplicates or exceptions without event logs.

Table 22: **Reference-first sampling yield, reconstructed.** Attempts are processed reference draws; nodes are distinct subset/node pairs visited by those draws. Initial pass concerns only the reference weight, legal-action count and already-retained-node tests. Retention is retained samples divided by attempts, not the fraction of all public states admitting a useful contrast.

Scenario / seat	Attempts	Nodes	Initial pass	Retained	Retention (%)
BTN vs BB 3B IP	79	77	79	10	12.66
BTN vs BB 3B OOP	105	104	104	10	9.52
BTN vs BB SRP IP	115	114	115	10	8.70
BTN vs BB SRP OOP	82	81	81	10	12.20
BTN vs SB 3B IP	54	52	54	10	18.52
BTN vs SB 3B OOP	126	121	126	10	7.94
SB vs BB 3B IP	135	132	135	10	7.41
SB vs BB 3B OOP	114	114	114	10	8.77
SB vs BB SRP IP	92	90	92	10	10.87
SB vs BB SRP OOP	81	80	81	10	12.35
Total	983	965	981	100	10.17

Indices are uniform over stored hand rows within a subset, not public nodes or reach-weighted play. The 10.17% yield describes this target-filtered, quota-stopped construction stream; it does not estimate all-state coverage or training holdout.

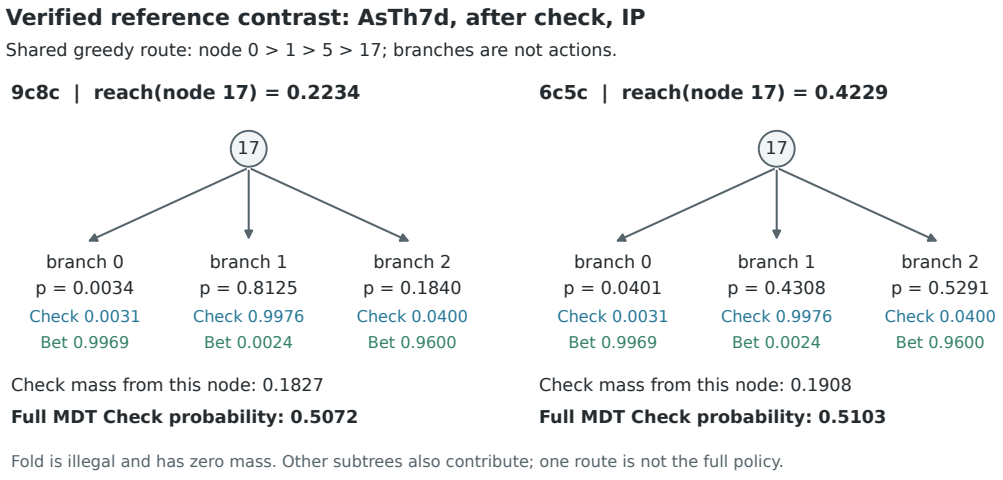

Figure 5: **Numerical routing contrast supplied to the LLM.** The example is the first case by identifier in Appendix G.3, paired with its first divergent shadow and selected independently of prediction outcomes. At node 17, the greedy branches differ, but both full policies remain mixed. Leaves retain soft action distributions; the action mask is applied to leaves, not branches. Numbers are rounded from the verified forward computation.

All effective inputs at the displayed router. The following five channels are the actual active inputs of node 17, not the largest weights of a dense router. $\delta_j = h_{17,j,1}(x_j) - h_{17,j,2}(x_j)$ is a raw branch-score difference, not an action probability or causal feature effect. No target route is shown.

Column and quantity	$x(9c8c)$	$x(6c5c)$	$\delta_j(9c8c)$	$\delta_j(6c5c)$
121 Provider unblocker score	5.0000	5.0000	4.4739	4.4739
122 Blocked-opponent equity	0.3895	0.3304	1.8450	0.4661
124 Mean own-card frequency	0.0510	0.0301	2.2098	0.3421
139 Equity minus range mean	-0.1480	-0.3895	0.9158	-5.9453
146 Conditional EV difference	0.2813	0.2007	-0.0012	-1.1949
Sum of raw score differences			9.4432	-1.8581

Column 121 is an opaque provider score, not a calibrated probability or an independently verified blocker measure. Column 122 is the range-weighted mean equity of opponent hands sharing exactly one private card, ignoring weights $\leq 10^{-6}$ and returning zero if blocked mass is below 10^{-6}; it is not an equity change. Column 124 averages the two cards' own-range inclusion frequencies, not inverse scarcity. Column 139 subtracts the acting range's weighted mean equity from hand equity. Column 146 is $(\mathrm{EV}_{\text{opp call}} - \max \mathrm{EV}_{\text{opp raise}})/\mathrm{pot}$, an oracle-conditioned continuation difference, not a same-node action-value gap.

Probability accounting. For each displayed reference, the node contribution to Check is $r_{17} \sum_{b=0}^{2} p_{17,b} q_b(\mathrm{Check})$. For 9c8c, this is 0.182697; the other subtrees contribute 0.324466, giving 0.507163. For 6c5c, the corresponding values are 0.190773, 0.319531, and 0.510304. Full-precision reconstruction checks normalization, legal-action masks, every leaf mass and the complete mixture.

E.4 Prompt Inputs, Execution and Representative Display

Inputs and response format. The cross-model study sends one structured prompt directly to each fixed predictor. Direct supplies context and target hand; Summaries adds named numerical inputs; SCCS supplies four references, displayed paths/impacts, model/solver policies and target summaries. It prioritizes nearby solver patterns over MDT predictions and requests JSON probabilities with an empty analysis field. No separate articulation call or scored natural-language rule is required.

Direct, Direct+Summaries and SCCS comprise 300 target-condition prompts on 100 targets. The response collection contains 2,723 raw and 2,723 scored records, including repeated attempts and reused responses; these totals are not counts of unique API requests. Greedy Route-only is evaluated separately, with 800 raw and 800 scored records. Three GPT configurations use the base prompt, while five DeepSeek/Gemini configurations append a JSON-output constraint. Execution settings and response selection are specified below.

Cross-condition input consistency. Across all 100 targets (50 IP and 50 OOP), 679 distinct target/feature pairs occur in both Direct+Summaries and the SCCS target block. Accounting for four-decimal versus two-decimal rounding, 674 pairs agree and five have opposite signs, spanning four IP targets: three `Hero EQ Adv` and two `Hero EV Adv` comparisons. For the fixed example below, the summaries are -0.1430 and -0.2859, whereas SCCS prints $+0.14$ and $+0.29$. The discrepancy arises from inconsistent display orientation: the summaries builder prints the stored OOP-minus-IP value, while the path renderer negates designated range-difference features for IP. Both use Hero-minus-Villain descriptions. This comparison covers common displayed target fields. Cross-model inputs and outputs are reported as scored; the matched-input evaluation fixes summary orientation across conditions. The cross-model comparison therefore includes numerical-input differences as well as changes in evidence organization.

Signed vulnerability proxy. In the cross-model prompts, `Hand Vuln` is described as the probability of being ahead and losing by the river. Its numerical value is the signed difference $V = A - E$, where A is its current-ahead score and E its current equity. This is not an outdraw-event probability and can be negative. Among the 100 Direct+Summaries prompts, 48 display this field and 28 of those values are negative; the fixed example prints -0.2408. A separate check of the 50 pilot input triples in Appendix B.1 agrees with $A - E$ within 2.99×10^{-8}, including 21 negative values. This check identifies the implemented formula; the cross-model outputs retain the prompt descriptions used at inference. The table below specifies the quantities used in forward-verified displays.

Execution settings. Table 23 separates returned response metadata from Greedy Route-only configurations. GPT-5.4 and GPT-5.5 returned dated model identifiers `gpt-5.4-2026-03-05` and `gpt-5.5-2026-04-23`; their successful cross-model responses report temperature 1 and strict JSON-schema output. A returned value does not imply that it was explicitly supplied rather than defaulted by the provider. The DeepSeek/Gemini response records do not specify all decoding settings; Greedy Route-only settings do not establish the settings of those other conditions.

Table 23: **Execution settings in the cross-model comparison.** Response metadata covers Direct, Direct+Summaries, and SCCS. A recorded output cap is not realized token usage. "Not recovered" is not a claim that the provider used no limit.

Configuration	Response metadata	Greedy Route-only configuration
GPT-5.4 medium	Responses API; medium; cap 3,600	Codex CLI 0.145.0; medium
GPT-5.5 low	Responses API; low; cap 3,600	Codex CLI 0.145.0; low
GPT-5.5 high	Responses API; high; caps 3,600 and 12,000	Codex CLI 0.145.0; high
DeepSeek Flash/Pro	Chat completions; full settings not recovered	Temperature 0; cap 3,600; JSON mode; thinking field omitted
Gemini Flash/Pro low/high	Native API; full settings not recovered	Matching low/high thinking; cap 3,600; temperature omitted

For GPT-5.5 high, 153 stored successful responses report a 3,600-token cap and 68 report 12,000. Sixty-eight prompt identities occur with both returned settings. Response matching uses model names, prompt ID and text, without separating temperature, token cap, reasoning budget or retry policy. These results therefore combine decoding budgets and repeated attempts. Appendix F.4 specifies response selection, missingness and the cross-wrapper GPT comparison.

Representative inputs. The first cross-model evaluation case by identifier is `QsTs` on `Ks6h2d`, `BTN_vs_BB_3B`, after a check with IP acting. Its conditions use the display conventions specified above. Illustrative cases outside this evaluation set do not contribute to aggregate errors. The paired example below instead shows the forward-verified input used for the compact-reference comparison.

Definitions for forward-verified displays. The table specifies the retained quantities, their units and orientations, including provider-defined scores and source-derived proxies.

Quantity	Correct interpretation
Latent branch probability	Normalized-softplus routing, with no action-legality mask. Branches are not actions.
Action/leaf mass	Legal-action softmax at leaves; sum probability-weighted leaf distributions over all routes.
Historical ahead score	Ordinal made-category comparison with half credit for equal categories; not exact hand-versus-range ahead probability.
Vulnerability	Signed historical ahead score minus equity, not the probability of being outdrawn.
Physical "kicker"	Lower private-card rank divided by 14; neither the higher private rank nor best-five-card kicker strength.
Range differences	Stored OOP-minus-IP quantities keep that orientation; conversion to acting-minus-opponent must be explicit and consistent.
History "MDF" proxy	Retained history-derived ratio; not the break-even equity of a call, especially when a token denotes a raise-to amount.
Card "scarcity"	Own-range card inclusion frequency, not its inverse.
Conditional "regret"	Response-line EV differences, not the target's same-node action-value gaps.
Provider blocker scores	Named provider scores with unrecovered formulas, not independently verified probabilities or causal effects.

E.4.1 Forward-Verified Paired Input

The first ID in Appendix G.3 targets `5h4h` on `AsTh7d` after a check, IP in `BTN_vs_BB_3B`. Both conditions share all five numerical input rows and four reference labels; only treatment adds

reference contrasts. Below, the full-precision rows and the first divergent pair are shown; the other two contrast records and repeated input definitions are omitted. Source-defined field meanings appear in Appendix D.4; this compact-contrast test uses its own input set, including fields withheld in the full-reference-path test. Probability accounting appears in Appendix E.3, with Figure 5. This excerpt illustrates the structured evidence supplied to the model; it is not a generated rationale.

```
Public state: {"Board":"AsTh7d","History":"k","Player":"IP","Pot":26.5,"Scenario":"
    BTN_vs_BB_3B"}
Target hand: 5h4h
Slots: check_call, fold, bet_raise; mask: [1.0, 0.0, 1.0]
Legal labels: Check, Bet(0.48x). Fold is not legal.

Feature IDs/formulas: see the common definitions and router table.

IDENTICAL FEATURE VALUES IN BOTH CONDITIONS
5h4h:
  121:hand_unblocker_score = 3.0
  122:block_opp_eq = 0.4361107647418976
  124:card_scarcity_avg = 0.02096233144402504
  139:rel_eq_diff_avg = -0.35072582960128784
  146:regret_call_vs_raise = 0.2940644919872284
9c8c:
  121:hand_unblocker_score = 5.0
  122:block_opp_eq = 0.38948071002960205
  124:card_scarcity_avg = 0.0510190911591053
  139:rel_eq_diff_avg = -0.14796501398086548
  146:regret_call_vs_raise = 0.2812754809856415
  Solver=[0.10364185273647308, 0.0, 0.8963581919670105]; MDT=[0.5071629285812378, 0.0,
     0.49283698201179504]
KsJd:
  121:hand_unblocker_score = 7.0
  122:block_opp_eq = 0.4786164462566376
  124:card_scarcity_avg = 0.06746646016836166
  139:rel_eq_diff_avg = -0.12745362520217896
  146:regret_call_vs_raise = 0.5448769330978394
  Solver=[0.7549523115158081, 0.0, 0.2450477033853531]; MDT=[0.5136809945106506, 0.0,
     0.48631876707077026]
6c5c:
  121:hand_unblocker_score = 5.0
  122:block_opp_eq = 0.3304012417793274
  124:card_scarcity_avg = 0.030078258365392685
  139:rel_eq_diff_avg = -0.3895149230957031
  146:regret_call_vs_raise = 0.20066848397254944
  Solver=[0.004403342958539724, 0.0, 0.9955966472625732]; MDT=[0.5103039741516113, 0.0,
     0.48969602584838867]
5s4s:
  121:hand_unblocker_score = 3.0
  122:block_opp_eq = 0.2255893051624298
  124:card_scarcity_avg = 0.012921396642923355
  139:rel_eq_diff_avg = -0.35223984718322754
  146:regret_call_vs_raise = 0.3139839768409729
  Solver=[0.222207173705101, 0.0, 0.7777928113937378]; MDT=[0.54618901014328, 0.0,
     0.45381075143814087]

TREATMENT ONLY: REFERENCE-ONLY CONTRASTS
Branches are latent. Score differences are raw scores, not probabilities.
A path or node is only part of the full action mixture.
Router 17, depth 3, first divergence=True
Compared branch scores: [1, 2]
  Reference 0, reach=0.22336577
  Branch probabilities: [0.0034459526650607586, 0.8125358819961548, 0.18401813507080078]
  Branch-conditional action mixtures:
    Branch 0: [0.003110976191237569, 0.0, 0.996889054775238]
    Branch 1: [0.9975722432136536, 0.0, 0.0024277002085000277]
    Branch 2: [0.03997056931257248, 0.0, 0.9600294828414917]
  Feature score differences:
    121:hand_unblocker_score = 4.4738922119140625
    122:block_opp_eq = 1.8449959754943848
    124:card_scarcity_avg = 2.209766387939453
    139:rel_eq_diff_avg = 0.9157689809799194
    146:regret_call_vs_raise = -0.001174628734588623
  Reference 2, reach=0.42294851
  Branch probabilities: [0.04009650647640228, 0.43082788586616516, 0.529075562953949]
  Branch-conditional action mixtures:
    Branch 0: [0.003110976191237569, 0.0, 0.996889054775238]
    Branch 1: [0.9975722432136536, 0.0, 0.0024277002085000277]
    Branch 2: [0.03997056931257248, 0.0, 0.9600294828414917]
  Feature score differences:
    121:hand_unblocker_score = 4.4738922119140625
    122:block_opp_eq = 0.4661250412464142
```

```
    124:card_scarcity_avg = 0.3421264588832855
    139:rel_eq_diff_avg = -5.945294380187988
    146:regret_call_vs_raise = -1.1949154138565063
Other reference contrasts omitted from this representative display.

Target solver policy, target MDT policy and target route are not supplied.
Return only finite, nonnegative JSON probabilities summing to one.
No natural-language rationale is claimed or scored.
```

F Fixed-LLM Strategy Communication with Reference Packages

F.1 Target-First Packages on Held-Out Boards

This experiment tests the complete SCCS reference package on queries fixed without consulting their solver policies. It uses the remaining 320 nodes from the grouped test partition in Appendix C.1, complementing the 320 nodes in Appendix G.1. Each of the same 64 canonical board groups contributes five nodes; each of the ten scenario/seat subsets contributes 32. The primary-seed Top-5 MDTs are reused without further training. These boards were excluded from MDT training and validation.

Queries and reference selection. At each node, the fifth presampled positive-reach hand is the query, and the first four define the uniform-reference control. The first hand also anchors SCCS. The query is excluded from every reference pool. SCCS searches the node's full positive-reach hand pool with the original MDT-policy distance thresholds (0.05 from the anchor and 0.03 between shadows), first-route divergence, and raw-feature ranking. It finds three shadows at 298 nodes. At the remaining 22, a fixed fallback uses the four uniform references; all 320 queries remain in the analysis. A second control uses the four nearest hands by mean absolute feature distance, dividing each valid, nonconstant feature by its span across the node's hand pool. Query identities and all reference sets are frozen before solver labels are extracted. Neither target policy distance nor mixing is used to retain queries.

Five inputs. Direct provides public context, the target hand, and legal action slots. Direct+Summaries adds the same named global set of 50 MDT input summaries at each node. Uniform, Nearest, and SCCS additionally supply four reference hands, their summaries, solver and MDT policies, and three compact reference-route contrasts. All three reference conditions use the same forward-verified renderer and input definitions. Summaries include oracle-derived equity and action-value quantities. The target solver policy, target MDT policy, and target routing and leaf information are absent from every input. Replacing either private target policy leaves all prompts unchanged; separate tests check invariance to target routing and leaf tensors. Independent reconstruction agrees with the MDT forward computation to maximum absolute error 4.71×10^{-7}. Under the retained tokenizer, median input lengths are 276, 3,399, 8,477, 8,495.5, and 8,492.5 tokens for Direct, Direct+Summaries, Uniform, Nearest, and SCCS, respectively. Reference conditions share a format and reference count but are not padded to exact token equality. These local counts exclude the provider's additional interface context.

Inference and analysis. DeepSeek-Flash and `gpt-6-sol` use the same configurations as in Appendix G.1, with 1,600 single-attempt inputs per model. Every query is scored. Invalid or missing outputs receive the fixed legal-uniform fallback; accepted probabilities sum to one within 10^{-3} and are renormalized. Four planned paired contrasts compare SCCS with Direct, Direct+Summaries, Uniform, and Nearest. We resample the 64 whole board groups with 50,000 bootstrap draws and report 98.75% intervals, adjusting for four comparisons within each model. Common-valid and selection-success subsets are secondary summaries. Mean-label, inverse-feature-distance, and routing-overlap interpolation use the same reference labels as the LLM; the withheld complete MDT is a separate numerical reference point. Feature-distance weights use the full valid feature pool. Routing weights additionally access the target's undisplayed MDT routing, so their inputs exceed the evidence available to the LLM in this experiment.

Results. Table 24 includes all 320 queries for each input. All 1,600 GPT responses use isolated inference sessions and satisfy the probability checks. DeepSeek returns all 1,600 responses, with one

invalid SCCS output receiving the predefined fallback. An independent implementation reproduces every case error, validity decision, numerical baseline, and adjusted interval.

Table 24: Target-first package comparison on all 320 queries. Lower summed action L_1 is better. Invalid outputs receive the predefined legal-uniform fallback.

Input	DeepSeek L_1	Valid	GPT L_1	Valid
Direct	0.8538	320/320	0.5093	320/320
Direct+Summaries	0.7646	320/320	0.4928	320/320
Uniform references	0.5611	320/320	0.4207	320/320
Nearest references	0.2962	320/320	0.0723	320/320
SCCS	0.6305	319/320	0.4035	320/320

SCCS lowers error relative to Direct by 26.2% for DeepSeek and 20.8% for GPT. Both improvements, and both comparisons with Direct+Summaries, meet the adjusted criterion in Table 25. These estimates measure the reference-package benefit on independently fixed queries. Uniform-reference contrasts include zero for both models, while nearest-feature references yield lower error than SCCS. All three reference conditions share the labeled-reference and route-contrast format; their comparison varies reference construction within that interface.

Table 25: All four prespecified contrasts per model: SCCS minus comparator, with negative values favoring SCCS. The 98.75% intervals use 50,000 board-cluster resamples and adjust for four comparisons within each model.

Model	Comparator	Difference	Adjusted interval
DeepSeek	Uniform references	0.0695	$[-0.0244, 0.1626]$
DeepSeek	Nearest references	0.3343	$[0.2328, 0.4387]$
DeepSeek	Direct+Summaries	-0.1340	$[-0.2484, -0.0218]$
DeepSeek	Direct	-0.2233	$[-0.3443, -0.1066]$
GPT	Uniform references	-0.0173	$[-0.0780, 0.0415]$
GPT	Nearest references	0.3312	$[0.2716, 0.3932]$
GPT	Direct+Summaries	-0.0893	$[-0.1547, -0.0254]$
GPT	Direct	-0.1059	$[-0.1791, -0.0328]$

On the 298 successful SCCS selections, its mean error is 0.6713 for DeepSeek and 0.4292 for GPT; uniform references yield 0.5967 and 0.4476, and nearest references yield 0.3121 and 0.0747. The corresponding 22-query fallback means are 0.0789 and 0.0554 for SCCS. These descriptive subsets preserve the main reference ordering. Restricting to jointly valid outputs also preserves all four adjusted conclusions for each model.

Table 26: Numerical label-transfer baselines on the same 320 queries. Feature-distance weights use the full valid feature pool; routing weights use target routing absent from the LLM prompts. These are not input-matched prompt decoders.

References	Label mean	Feature distance	Routing overlap
Uniform references	0.6231	0.5616	0.5146
Nearest references	0.1611	0.0690	0.1571
SCCS	0.6467	0.6203	0.6089

The private MDT output has summed L_1 of 0.4254; it is not supplied to the LLM. The numerical predictors preserve the reference ordering in the LLM comparison (Table 26), with the additional inputs specified above. Both evaluations use solver-derived summaries.

Contrast generation and target-first retrieval. SCCS selects policy-divergent reference hands around an anchor, whereas the nearest-reference control selects hands similar to an already fixed query. The former emphasizes changes in policy across hands; the latter supplies nearby labeled examples for prediction. Tables 24 and 25 measure query accuracy under these constructions. SCCS defines the policy-contrast condition used in the communication evaluation, alongside the reported retrieval controls.

F.2 Cross-Model Errors and Coverage

Table 27: **Parsed-output coverage in the cross-model study.** Each condition has 100 selected targets. Counts follow the response-selection rule described in Appendix F.4; they are not first-attempt API success rates.

LLM/API configuration	Direct	+Summaries	Greedy Route-only	SCCS
Gemini-3.1 Flash	99	89	99	100
Gemini-3.1 Pro low	100	100	100	100
Gemini-3.1 Pro high	22	20	100	15
DeepSeek-V4 Flash	100	100	100	100
DeepSeek-V4 Pro	100	100	100	99
GPT-5.4	99	99	100	68
GPT-5.5 low	100	100	100	100
GPT-5.5 high	36	42	100	29

F.3 Cross-Model Aggregation and Input Conventions

The eight-configuration comparison uses the rendering, input orientations and wrappers documented in Appendices B.1 and E.4. Coverage and retention are in Appendix F.4; full errors in Appendix F.2. Verified DeepSeek/GPT comparisons retain their own inputs, populations and estimates.

Within this heterogeneous, parse-conditional protocol, SCCS reduces mean action-averaged solver L_1 from 0.211 to 0.100 versus Direct and from 0.173 to 0.100 versus Greedy Route-only. The separately reported highest-probability action agreement rises from 57.2% to 76.1% over seven configurations, excluding Gemini Pro high; it is not the eight-configuration aggregation used for L_1. The full-precision Gemini Pro high SCCS mean is 0.0897083, rounded to 0.090 in the table. Direct+Summaries has mean error 0.256. Comparing its inputs with SCCS reveals five opposite-sign common-feature comparisons across four targets, alongside other input differences. The forward-verified study holds target feature values and definitions constant across feature-bearing conditions and reports its results separately.

F.4 Paired Cross-Model and Reference-Only Diagnostics

Inputs and aggregation. Greedy Route-only receives the current prediction target's own greedy trace: visited nodes, branch/reach probabilities, summary impacts and a leaf prototype. The requested output is that same hand's full solver strategy. No other hand's labeled policy, target solver policy or aggregate MDT mixture is supplied. SCCS instead provides other hands' reference labels, comparisons and display structure to predict the target. The SCCS communication task withholds the target MDT output: a complete target computation would directly determine that model's prediction, whereas the reference-based task requires predicting another hand from the supplied comparisons. The target-trace baseline and the reference-path ablations test different input conditions within this evaluation. Both use the renderer described in Appendix B.1. Direct+Summaries supplies unstructured target summaries. The cross-model comparison measures package differences; reference-selection and matched-path controls are reported separately.

Table 3 averages each condition's parse-success cases, then the eight configuration means equally. Table 29 also restricts Greedy Route-only/SCCS to identical parsed targets within each configuration.

Coverage is in Table 27: Greedy Route-only has 100 parsed targets in seven configurations and 99 for Gemini Flash. The GPT Greedy Route-only calls use a separate wrapper from SCCS despite matching named models/reasoning settings.

Missingness and uncertainty. Canonical SCCS failures are 85 `api_error` for Gemini Pro high; 49 `api_error` and 22 `no_json` for GPT-5.5 high; 32 `no_json` for GPT-5.4; and one `no_json` for DeepSeek Pro. These labels do not reveal provider causes or unobserved prediction quality. Response selection prioritizes parsed records and then recency. This coverage is not first-attempt reliability.

Table 28: **Fixed-LLM strategy prediction on targets absent from the displayed references.** Lower L_1 is better. Each model entry reports mean $\pm$ standard error over parsed target cases. The Mean row reports mean $\pm$ sample standard deviation across eight LLM configurations.

Solver target

LLM configuration	Direct	+Summaries	Greedy Route-only	SCCS
Gemini-3.1 Flash	0.211 ± 0.015	0.223 ± 0.016	0.144 ± 0.010	0.099 ± 0.009
Gemini-3.1 Pro low	0.235 ± 0.020	0.299 ± 0.021	0.228 ± 0.015	0.106 ± 0.010
Gemini-3.1 Pro high	0.268 ± 0.019	0.303 ± 0.015	0.205 ± 0.015	0.090 ± 0.012
DeepSeek-V4 Flash	0.185 ± 0.015	0.263 ± 0.019	0.212 ± 0.014	0.116 ± 0.013
DeepSeek-V4 Pro	0.254 ± 0.016	0.255 ± 0.017	0.174 ± 0.011	0.112 ± 0.011
GPT-5.4	0.187 ± 0.013	0.216 ± 0.014	0.124 ± 0.009	0.095 ± 0.010
GPT-5.5 low	0.158 ± 0.013	0.241 ± 0.016	0.150 ± 0.010	0.096 ± 0.009
GPT-5.5 high	0.186 ± 0.007	0.248 ± 0.017	0.147 ± 0.009	0.086 ± 0.010
Mean	0.211 ± 0.038	0.256 ± 0.032	0.173 ± 0.038	$\mathbf{0.100 \pm 0.011}$

MDT target

LLM configuration	Direct	+Summaries	Greedy Route-only	SCCS
Gemini-3.1 Flash	0.204 ± 0.014	0.224 ± 0.015	0.142 ± 0.009	0.114 ± 0.009
Gemini-3.1 Pro low	0.235 ± 0.020	0.300 ± 0.020	0.234 ± 0.014	0.124 ± 0.010
Gemini-3.1 Pro high	0.261 ± 0.019	0.314 ± 0.012	0.209 ± 0.015	0.109 ± 0.013
DeepSeek-V4 Flash	0.173 ± 0.013	0.256 ± 0.018	0.207 ± 0.013	0.130 ± 0.012
DeepSeek-V4 Pro	0.252 ± 0.014	0.259 ± 0.014	0.171 ± 0.010	0.125 ± 0.010
GPT-5.4	0.175 ± 0.011	0.210 ± 0.014	0.119 ± 0.008	0.104 ± 0.008
GPT-5.5 low	0.155 ± 0.011	0.241 ± 0.016	0.149 ± 0.009	0.109 ± 0.009
GPT-5.5 high	0.173 ± 0.008	0.260 ± 0.015	0.147 ± 0.008	0.097 ± 0.008
Mean	0.204 ± 0.041	0.258 ± 0.035	0.172 ± 0.040	$\mathbf{0.114 \pm 0.011}$

Table 29: **Greedy Route-only versus SCCS under two aggregation conventions.** Values are action-averaged L_1; parenthesized values are relative reductions from Greedy Route-only to SCCS.

Aggregation	L_1 **to solver**	L_1 **to MDT**
Condition-wise Table 3 means	$0.173 \to 0.100$ (42.2%)	$0.172 \to 0.114$ (33.7%)
Shared-parse paired means	$0.170 \to 0.100$ (41.2%)	$0.172 \to 0.114$ (33.8%)

The post-hoc paired analysis resamples the 100 public-node targets 10,000 times (seed 20260917), using the same draws across configurations and each configuration's valid pairs. Equal-configuration mean difference is -0.07005, 95% interval $[-0.09382, -0.04598]$. Per-configuration intervals are unadjusted; repeated model observations of one target are not independent game states. Selection and wrapper differences remain.

Table 30: **Cross-model paired uncertainty and coverage sensitivity.** Δ is SCCS minus Greedy Route-only mean action L_1 on common parsed targets. R/S give Greedy Route-only/SCCS coverage out of 100. The final column is all-target error of the post-hoc SCCS-to-Route fallback, not SCCS alone.

Configuration	R/S	Δ	95% interval	Fallback
Gemini Flash	99/100	$-.0452$	$[-.0707, -.0198]$	.0994
Gemini Pro low	100/100	$-.1219$	$[-.1551, -.0899]$	.1058
Gemini Pro high	100/15	$-.0950$	$[-.1981, .0024]$	.1910
DeepSeek Flash	100/100	$-.0952$	$[-.1302, -.0607]$	.1163
DeepSeek Pro	100/99	$-.0641$	$[-.0928, -.0350]$	.1106
GPT-5.4	100/68	$-.0329$	$[-.0650, .0007]$	.1018
GPT-5.5 low	100/100	$-.0544$	$[-.0796, -.0281]$	.0958
GPT-5.5 high	100/29	$-.0516$	$[-.0888, -.0114]$	.1319

The post-hoc fallback uses SCCS if parsed, otherwise Greedy Route-only, otherwise legal-uniform; the comparator is Greedy Route-only with the same uniform fallback. It covers all 100 targets per configuration: mean 0.11908 versus 0.17294, difference -0.05385, target-bootstrap 95% interval $[-0.06920, -0.03791]$. This composition does not impute SCCS's missing answers or measure live fallback cost. Gemini Pro low and DeepSeek Flash have complete paired coverage under their respective wrappers.

F.4.1 Reference-only baseline diagnostic

Two deterministic predictors use only the four solver policies printed to two decimals: their mean, or the L_1 medoid minimizing total reference distance. Illegal slots are masked and vectors normalized; ties use display order and 10^{-12} tolerance. Neither reads target features or target policy. Action-averaged errors on all 100 targets are 0.12459 (mean), 0.14923 (medoid), 0.16187 (legal-uniform) and 0.05900 (private target MDT). Full-precision reference labels give mean/medoid $0.12447/0.14666$; the primary comparison uses printed information. No model calls or fitting are involved.

Table 31: **Reference-only prediction on SCCS's parsed targets.** M is the mean, C the medoid, and S the SCCS answer. Intervals for S–M and S–C are paired 10,000-target-bootstrap percentile intervals, seed 20260917, unadjusted across configurations. Negative differences favor SCCS.

Configuration	N	M	C	S	95% S–M	95% S–C
Gemini Flash	100	.1246	.1492	.0994	$[-.0496, -.0007]$	$[-.0805, -.0221]$
Gemini Pro low	100	.1246	.1492	.1058	$[-.0447, .0074]$	$[-.0749, -.0149]$
Gemini Pro high	15	.1179	.1551	.0897	$[-.0954, .0491]$	$[-.1639, .0280]$
DeepSeek Flash	100	.1246	.1492	.1163	$[-.0319, .0167]$	$[-.0602, -.0065]$
DeepSeek Pro	99	.1238	.1486	.1117	$[-.0388, .0150]$	$[-.0698, -.0069]$
GPT-5.4	68	.1213	.1371	.0947	$[-.0536, .0011]$	$[-.0767, -.0125]$
GPT-5.5 low	100	.1246	.1492	.0958	$[-.0535, -.0035]$	$[-.0844, -.0248]$
GPT-5.5 high	29	.1184	.1466	.0861	$[-.0804, .0201]$	$[-.1274, -.0027]$

Dependence on reference proximity. The nearest-reference distance uses the hidden target policy and is therefore an explanatory stratum, not an executable retrieval rule. The four fixed intervals $[0, .025)$, $[.025, .05)$, $[.05, .10)$, and $[.10, \infty)$ contain 32, 29, 31, and 8 targets. Reference-mean errors are respectively 0.1151, 0.1425, 0.1263, and 0.0911. The two configurations with complete coverage in both provider conditions give SCCS errors $0.0656/0.0922/0.1407/0.1800$ (Gemini Pro low) and $0.0704/0.0936/0.1726/0.1645$ (DeepSeek Flash). Their advantage over the mean is confined to the closer two strata. These post-hoc strata describe the policy-filtered population, with few targets in the last interval. The independently fixed-query comparison is reported in Appendix F.1 and the information-matched path test in Appendix G.1.

All eight SCCS point estimates beat the predefined reference-only predictors on matched parsed cases; two unadjusted intervals against the mean exclude zero. These controls apply numerical prediction rules to the reference labels.

Holding the dominant action fixed. We perform two post-hoc, parse-conditional diagnostics. First, among 342 matched model–target cases where Direct already selects the solver argmax, Greedy Route-only and SCCS have nearly identical argmax agreement (81.3% and 80.1%), but SCCS lowers Greedy Route-only solver L_1 from 0.15633 to 0.10082, a 35.5% reduction. SCCS also lowers Direct's remaining solver L_1 from 0.11948 to 0.10082, a 15.6% reduction.

Second, we restrict further to the 230 model–target cases where Direct, Greedy Route-only, and SCCS all select the correct solver argmax. SCCS lowers Greedy Route-only solver L_1 from 0.12506 to 0.07648 (38.8%) and MDT L_1 from 0.14908 to 0.09984 (33.0%). The configuration-level point estimate favors SCCS in all eight configurations under both targets. Because dominant-action correctness is fixed, these differences measure the remaining probability allocation rather than an argmax-action correction. They are diagnostic subset analyses, not a formal probability-calibration study, and repeated target states across model configurations are not treated as independent game samples.

F.5 Matched Inputs on Reference-First Targets

To compare evidence packages with consistent numerical definitions, this experiment fixes all 100 reference-first targets, four-reference identities and models, and compares forward-verified inputs under one DeepSeek-V4 Pro protocol. The population uses target-label selection and may overlap fitting.

Four inputs. Direct supplies context, hand and legal actions, with neither MDT evidence nor reference labels; it is a package comparator, not a matched causal ablation. Summaries adds target inputs: the union of the 40 target summaries used in the cross-model comparison and effective inputs needed by the displays. Route adds the prediction target's own four greedy-route nodes, effective inputs, raw terms, branch probabilities and the reached leaf prototype, without references or complete target MDT policy. SCCS instead adds four reference solver/MDT policies and three compact reference contrasts, without target route or full MDT policy. Summary-bearing arms have byte-identical target values and definitions. Changing the withheld target solver answer leaves the prompt unchanged. These are oracle-assisted evidence packages; the reference addition bundles labels and contrasts.

Execution and analysis. The official wrapper uses temperature zero, thinking disabled, JSON and a 128-token cap. All 400 requests are frozen and single-attempt, without replacement or interim effect inspection. Seed 20260923 shuffles cases and arm assignments: each of 24 permutations occurs four times, with four extra orders. Model aliases are not pinned weight revisions. Forward reconstruction agrees within 2×10^{-6} with fixed weights; 18 tests check input consistency, probability validity and independent scoring. No output reaches the cap; four invalid outputs receive fallback.

Errors sum three action slots. The parser requires exactly three finite, nonnegative legal probabilities summing within 10^{-6} of one and rejects duplicate keys. Missing/invalid outputs use legal-uniform fallback, retaining all targets. Primary Direct-minus-SCCS and Route-minus-SCCS differences use 50,000 paired, case-weighted bootstrap draws over 98 exact board groups, seed 20260923. Both Bonferroni 97.5% lower endpoints must exceed zero for package benefit; 95% intervals and subgroups are descriptive.

Table 32: **Matched-input results on reference-first targets.** Lower is better. All 100 targets remain in every arm. Input tokens are provider-reported totals.

Input	Valid	Summed L_1	Action-averaged L_1	Input tokens
Direct	100/100	0.516316	0.172105	30,249
Summaries	99/100	1.011069	0.337023	403,116
Verified target route	100/100	0.977355	0.325785	518,222
SCCS reference package	97/100	0.321596	0.107199	900,710

Table 33: **Paired package effects.** Positive favors SCCS.

Comparison	Mean	95% interval	Adjusted 97.5% interval
Direct minus SCCS	0.19472	$[0.08527, 0.30753]$	$[0.06999, 0.32345]$
Route minus SCCS	0.65576	$[0.56724, 0.74432]$	$[0.55501, 0.75689]$

Both adjusted intervals favor SCCS on this selected population. The comparison adds forward-verified evidence under a common wrapper; it does not re-estimate the original 52.6% result, isolate contrast utility beyond labels or certify verbal reasoning. Evidence lengths differ, and all failures remain under the fixed cap. Appendix G.4 isolates reference contrasts.

F.6 Cross-Model Tests with Matched Inputs

DeepSeek and GPT receive identical target identities, reference hands and input text in three matched comparisons. The package comparison uses all 100 targets and four conditions from Appendix F.5; the reference-selection comparison uses all 100 targets and two conditions from Appendix H.2;

Table 34: **All scenario/seat subgroups.** Ten targets per row; summed errors. These comparisons are descriptive, not separately powered tests.

Scenario/seat	Direct	Summaries	Route	SCCS
BTN_vs_BB_3B_IP	0.4348	0.9324	1.2469	0.3492
BTN_vs_BB_3B_OOP	0.4372	0.6570	0.8287	0.4250
BTN_vs_BB_SRP_IP	0.3606	1.0430	1.0593	0.2545
BTN_vs_BB_SRP_OOP	0.6731	1.3894	0.7358	0.1679
BTN_vs_SB_3B_IP	0.5412	0.9720	0.8831	0.2846
BTN_vs_SB_3B_OOP	0.4915	0.8357	0.9282	0.4507
SB_vs_BB_3B_IP	0.2165	1.1771	1.1585	0.3240
SB_vs_BB_3B_OOP	0.8159	1.2603	0.8932	0.4791
SB_vs_BB_SRP_IP	0.6083	0.9339	1.0034	0.2283
SB_vs_BB_SRP_OOP	0.5841	0.9099	1.0365	0.2526

the compact-path comparison uses all 300 targets and three conditions from Appendix G.7. Each comparison pairs GPT predictions with the DeepSeek-V4 Pro responses specified in its respective appendix; the three populations are analyzed separately.

Model interface. Each GPT prediction uses an isolated `gpt-6-sol` inference session with medium reasoning and receives only the fixed experimental input plus the common JSON-output instruction. No tools or external results are used. This interface differs from the GPT-5.4/5.5 endpoints in Table 3 and from DeepSeek chat completions; their decoding controls are not identical. The model identifiers describe the configurations used, not pinned provider weights.

Execution and scoring. The fixed GPT list comprises 400 package predictions, 200 reference-selection predictions and 900 compact-path predictions. Inputs are interleaved with the held-out experiment below using a fixed random seed before inference. Each input receives one attempt, without retries or replacement. Each isolated session uses the specified model and reasoning setting, without tools or additional user instructions; the platform's fixed environment metadata is common to all inputs. Responses that violate this protocol receive the same legal-uniform fallback as missing or invalid answers.

Each comparison uses its specified scoring rules. The package and reference-selection tests require probability sums within 10^{-6} and zero mass on illegal actions; the compact-path test accepts sums within 10^{-3} and renormalizes accepted legal probabilities. These rules apply equally to the DeepSeek and GPT outputs within a test. The primary metric is summed action-distribution L_1, including every target. Additional uncertainty estimates resample canonical suit-equivalent board groups with 50,000 paired bootstrap draws. The package, reference-selection and compact-path tests have 98, 98 and 300 groups, respectively. Intervals adjust for three, one and two planned contrasts within each model and test. These are sensitivity intervals reported alongside each test's primary analysis.

Table 35: **Cross-model comparison on identical inputs.** All-target summed action L_1, lower is better. DS denotes DeepSeek-V4 Pro; GPT denotes `gpt-6-sol`. Coverage counts valid responses before fallback. Blocks correspond to separate target populations and interventions.

Test	Input	DS L_1	GPT L_1	DS valid	GPT valid
Package	Direct	0.5163	0.5064	100/100	100/100
	Summaries	1.0111	0.4838	99/100	100/100
	Greedy Route-only	0.9774	0.4735	100/100	100/100
	SCCS	0.3216	0.2960	97/100	100/100
References	Selected	0.3653	0.3025	100/100	100/100
	Uniform	0.6842	0.5369	100/100	100/100
Compact paths	Common	0.7248	0.3761	299/300	300/300
	Faithful	0.6841	0.3915	299/300	300/300
	Shuffled	0.8671	0.4277	300/300	300/300

Table 36: **Paired cross-model contrasts.** Positive differences favor SCCS, selected references or faithful paths. Intervals resample canonical board groups with the adjustment described above; each experiment's primary intervals are reported in its respective subsection.

Test	Contrast	Model	Difference	Adjusted interval
Package	Direct − SCCS	DS	0.1947	$[0.0610, 0.3312]$
		GPT	0.2104	$[0.0992, 0.3314]$
	Summaries − SCCS	DS	0.6895	$[0.5586, 0.8171]$
		GPT	0.1878	$[0.0753, 0.3058]$
	Route − SCCS	DS	0.6558	$[0.5493, 0.7638]$
		GPT	0.1775	$[0.0800, 0.2766]$
References	Uniform − Selected	DS	0.3189	$[0.1990, 0.4384]$
		GPT	0.2344	$[0.1415, 0.3310]$
Compact paths	Common − Faithful	DS	0.0406	$[-0.0390, 0.1219]$
		GPT	-0.0154	$[-0.0584, 0.0274]$
	Shuffled − Faithful	DS	0.1830	$[0.0996, 0.2694]$
		GPT	0.0362	$[-0.0153, 0.0881]$

All 1,500 GPT responses satisfy the inference protocol and output checks. The SCCS package reduces GPT error by 41.6% relative to Direct and improves over both feature and route conditions (Table 36). Selected references also improve prediction on the original SCCS-filtered targets. These findings extend the package and conditional reference results beyond DeepSeek. The compact-path result is different: GPT's mean error is lowest with common evidence, and neither adjusted path contrast excludes zero. The same prompt-only numerical decoder obtains 0.4994, while the withheld complete MDT obtains 0.3033. GPT therefore also shows that this particular numerical decoder is not an upper bound on use of the shared evidence. Neither cross-model replication nor the conditional reference comparison establishes a path-only benefit.

Input-matched target display. A post-hoc analysis of these existing package responses compares Summaries with Greedy Route-only. All 100 prompt pairs have identical instructions and feature content; only the verified target-route field is added. Original scoring, fallbacks and target membership are unchanged. Summaries-minus-Route summed L_1 is 0.0337 for DeepSeek and 0.0103 for GPT. With 50,000 paired resamples of the 98 suit-canonical board groups (seed 20260925), case-weighted 95% intervals are $[-0.0788, 0.1467]$ and $[-0.0731, 0.0966]$. Bonferroni 97.5% intervals for these two contrasts are $[-0.0947, 0.1627]$ and $[-0.0856, 0.1085]$. This adjustment concerns only the two added contrasts, not the full history of exploratory tests. Both mean differences favor the display, but neither interval excludes zero. The treatment includes a reached-leaf prototype as well as route computations; the original target selection and possible fitting overlap remain.

F.7 Draw-Aware Prediction Example

The GPT-5.5 high visible-rationale diagnostic uses the same public state and target in all conditions. The target `Td9c` is an offsuit gutshot; Table 37 records its predicted mixtures in public-context action order and action-averaged solver L_1. Direct and Summaries assign most probability to Fold, citing weak out-of-position realization. With the folding and continuing reference hands, the response instead associates the target with a continuing T9 draw and recovers a Call/Bet mixture. This selected illustration documents response behavior; it does not measure semantic comprehension or live exploitability.

G Controlled Path-Display Comparisons

These experiments test the contribution of displayed paths by fixing the reference labels, numerical inputs and MDT predictions within each comparison. They cover reference-only and target-connected displays, common-evidence and length/assignment controls, and policy-filtered or independently sampled queries. Sampling, available information and inference budgets are specified for each test; estimates are not pooled across these conditions.

Table 37: **Strategy prediction for an out-of-position target absent from the references.** SCCS supplies same-context folding and continuing reference hands.

Public context	SB vs BB single-raised pot, board `8s6h5d`, history `b4,b11`, OOP decision, actions {Call, Fold, Bet($1.79\times$)}.
Contrastive trace	Fold prototype `Jc4c`: model $[0.02, 0.97, 0.00]$, solver $[0.00, 1.00, 0.00]$. Policy-divergent continuing prototypes, including `Qh9h` and `Jc7c`, route to Call/Bet mixtures with solver Fold $= 0$. The historical display associates the contrast with draw strength, hand EV, and range-relative equity.
Unseen target hand	`Td9c`. Key SCCS values: draw strength 0.30 vs average 0.13; hand EV 0.03; hand equity 0.32; range-relative equity -0.14.
Solver target	$[\text{Call} : 0.63,\ \text{Fold} : 0.00,\ \text{Bet} : 0.37]$.
GPT-5.5 direct	$[0.12, 0.88, 0.00]$, average L_1 to solver $= 0.585$.
GPT-5.5 + summaries	$[0.07, 0.91, 0.02]$, average L_1 to solver $= 0.605$.
GPT-5.5 + SCCS	$[0.59, 0.00, 0.41]$, average L_1 to solver $= 0.025$.

G.1 Information-Matched Held-Out Communication

This experiment evaluates the same soft-path interface on boards excluded from MDT fitting. It uses the primary-seed Top-5 models and data partition from Appendix C.1, without further training or model selection. All 64 canonical test-board groups are included. Identity-only sampling selects five scenario/seat subsets per group, balanced to 32 nodes for each of the ten subsets, yielding 320 targets. At each node, the first four presampled positive-reach hands supply references and the fifth is the query. Neither target policies nor model errors determine this selection.

Four matched inputs. All conditions contain identical public context, named target/reference summaries, definitions, four reference solver labels and four reference MDT predictions. Common evidence contains these fields alone. Faithful paths add the target and reference routing displays, retaining at least 95% of each hand's mass and reporting omitted mass and active input identities. Shuffled paths derange the reference routing assignments while preserving the target routing and the other evidence. The repetition control adds repeated public context without new information. Faithful, shuffled and repetition inputs have equal lengths under the retained tokenizer. All 320 corresponding DeepSeek prompt-token triples are also exactly equal. No target solver policy, target MDT action distribution or leaf action policy is sent to either model.

Group disjointness, legal actions and card legality are checked before inference. Independent reconstruction of the displayed soft paths agrees with the checkpoint computation to maximum absolute error 4.41×10^{-7}. The same retained-mass renderer as in Appendix G.7 is used; it is not tuned to the held-out answers. The numerical reference decoder uses all 81 path overlaps and the same four reference labels. The complete MDT prediction is scored separately and never supplied to the language models.

Models and inference. DeepSeek's official API receives the model identifier `deepseek-flash`, with thinking disabled, temperature zero, JSON output and a 128-token output cap. The API returns that same alias. GPT uses the isolated `gpt-6-sol` medium-reasoning interface described above. Each model has 1,280 fixed single-attempt inputs. Failures count, are not retried, and use the predeclared legal-uniform fallback. DeepSeek-V4 Pro and DeepSeek-Flash results are reported separately.

Paired analysis. We retain all 320 targets and report whole-distribution L_1 and valid-output coverage. Three contrasts compare faithful paths with common evidence, repeated context and shuffled paths. Bootstrap resampling operates on the 64 board groups, each containing five queries, preserving all paired conditions. The 50,000 draws use seed 20260925; within-model Bonferroni adjustment gives 98.333% intervals for the three contrasts. Results are reported by model, without pooling providers or selecting a favorable contrast. The numerical decoder and private MDT are separate reference points rather than additional LLM conditions.

Both providers return all 1,280 scheduled responses; all GPT sessions pass the isolation checks. Three DeepSeek outputs fail the probability checks and receive fallback, as reflected in Table 38. An

Table 38: **Communication on 320 training-held-out queries.** All-target summed action L_1 and valid-output coverage. DS denotes DeepSeek-Flash; GPT denotes `gpt-6-sol`. Reference methods are scored on exactly the same targets.

Input	DS L_1	GPT L_1	DS valid	GPT valid
Common evidence	0.6294	0.4152	318/320	320/320
Faithful paths	0.6700	0.4281	319/320	320/320
Shuffled paths	0.7224	0.4578	320/320	320/320
Repeated context	0.6317	0.4364	320/320	320/320
Numerical routing	0.5455		–	–
Private complete MDT	0.4369		–	–

Table 39: **All planned held-out path contrasts.** Positive differences favor faithful paths. Every 98.333% interval resamples the 64 canonical board groups and includes zero.

Model	Contrast	Difference	Adjusted interval
DeepSeek-Flash	Common − Faithful	-0.0406	$[-0.1050, 0.0280]$
	Repeated − Faithful	-0.0382	$[-0.0992, 0.0230]$
	Shuffled − Faithful	0.0525	$[-0.0020, 0.1085]$
GPT	Common − Faithful	-0.0129	$[-0.0511, 0.0267]$
	Repeated − Faithful	0.0083	$[-0.0266, 0.0438]$
	Shuffled − Faithful	0.0297	$[-0.0143, 0.0759]$

independent implementation reproduces point estimates, coverage and every adjusted interval. The faithful-path input has higher mean error than common evidence for both models. Its lower error than shuffled paths does not reach the adjusted criterion for either model. These intervals neither establish an incremental path benefit nor demonstrate equivalence between inputs.

The numerical decoder outperforms DeepSeek's faithful-path predictions but not GPT's. This component test measures the effect of adding paths to shared features and labels on training-held-out boards. The original selected-target comparison and its cross-model replication instead evaluate the complete evidence package.

G.2 Full Reference Paths on Independently Sampled Targets

Population and shared inputs. This exploratory study asks whether complete reference paths improve prediction beyond the same labels and numerical inputs. It fixes 20 targets before responses, using the seed-999 BTN-versus-BB three-bet IP MDT. Boards may overlap fitting. Within a fixed sampling partition, seed-20260917 SHA-256 ranks board groups after excluding boards already used in recorded LLM prompts; the lowest-hash node represents each of the first 20 groups. A separate identity hash ranks legal positive-weight hands: first is target, next four are references. Selection precedes feature/policy access, with no target-label filters or replacements. This exclusion covers recorded queries only. Input arrays, masks, solver labels and models are fixed before inference.

Both arms receive context, identities, reference solver/MDT policies and the 37 effective-channel features defined in Appendix D.4. Nineteen target fields are null; reference rows retain source-present values, including shared node quantities. Target solver policy, MDT policy and route are hidden. Mutation checks cover excluded target fields, not all information correlated through references. Treatment adds actual-forward router probabilities, effective inputs, all 81 leaf masses and legal leaf policies. In this runtime, 15 of 40 routers use fewer than five effective inputs (minimum two). All 80 reference mixtures reconstruct within 5.92×10^{-8} leaf-mass and 2.51×10^{-7} policy error, without changing saved indices or masks.

Execution and estimand. Official DeepSeek-V4 Pro uses thinking disabled, temperature zero, JSON output and a 512-token cap. Seeded pair order puts each arm first ten times. All 40 single-attempt requests return `stop` with the same model/fingerprint; the alias is not a pinned revision. There are no retries or outcome-dependent revisions. The primary error sums absolute differences over check/call, fold and bet/raise; dividing by three gives fixed-slot action-averaged L_1. Valid

outputs have exactly three probabilities in $[0, 1]$, sum within 10^{-6} of one and no illegal mass. Invalid outputs receive legal-uniform fallback without repair. All targets remain. Control minus paths is

Table 40: **Full-reference-path comparison on 20 targets.** Errors are summed distribution L_1, lower is better; they are not on the action-averaged scale of Table 3. The two LLM conditions share all non-path evidence.

Condition / deterministic baseline	Valid LLM outputs	Mean summed L_1
Shared labels, features and reference MDT policies	19/20	0.490426
Same evidence plus numerical reference paths	19/20	0.680607
Mean of four reference solver policies	—	0.649714
Uniform over legal actions	—	0.863415

-0.190181, with paired-board 95% percentile interval $[-0.469161, 0.035398]$ (20,000 draws, seed 20260917). The same target in both arms sums to 1.000017 and fails validation. On 19 valid pairs the difference is -0.200190. At tolerance 10^{-12}, two improve, six worsen and twelve tie; a nominal 5.55×10^{-17} difference counts as a tie. Among valid outputs, 11 control and 12 path predictions

Table 41: **Per-target full-reference-path results.** Both columns use summed L_1. Positive control-minus-path values favor paths. The starred row uses the same predefined fallback in both conditions. Cards use rank/suit codes.

Board	History	Target	Control	Paths	Control−paths
9s4h4d	b17.8	AdJd	1.1988	1.1988	0.0000
Ks7s7h	k	JsJd	0.1561	0.1561	0.0000
Qh6s3s	k	5s5d	0.1208	0.1208	0.0000
AsQhJs	k	7s7h	0.2951	0.5369	−0.2418
9h6s3s	k	JsJd	0.2312	1.7688	−1.5377
9s8h4s	b12.8	KhJc	0.0196	1.9971	−1.9776
8s6h3d	b17.8	8d8c	0.7566	0.0020	+0.7546
KsJs7h	k	TdTc	0.6989	0.6989	0.0000
Kh3s2s	b12.8	AcQh	0.1354	0.1354	0.0000
5h4s2s	b17.8	JhJd	1.9992	1.9996	−0.0004
AsJhTs	k	Ah9h	0.2906	0.0946	+0.1961
As2s2h	k	AdJh	0.0472	0.4899	−0.4427
KsKh3d	k	AcTc	0.5418	0.5418	0.0000
As5h4s*	b17.8	AdKh	1.3197	1.3197	0.0000
KsTh3d	b12.8	AcJc	0.1039	0.1039	0.0000
Ts9h3s	k	AhQc	0.0786	0.0786	0.0000
JsJh5d	k	7d6d	0.4904	0.4904	0.0000
Ts9s9h	b12.8	Ah8h	0.0198	0.0198	0.0000
6s4h2d	b17.8	Ad7d	0.3425	0.3425	0.0000
AsJsTh	k	KhQd	0.9624	1.5164	−0.5540

match a reference MDT policy within 10^{-6} per action. Input totals are 117,881 versus 485,839 tokens ($4.12\times$). The path-effect interval includes zero. This comparison uses one scenario/checkpoint, unequal prompt lengths, oracle reference evidence and possible training overlap. References are independently sampled rather than selected by SCCS.

G.3 Compact Reference Contrasts on Twenty Filtered Targets

Population and contrast. This comparison tests whether a concise contrast at a divergent router is useful beyond the shared reference evidence. Two IDs per reference-first scenario/seat subset are chosen by seed-20260920 identity hash: 20 distinct board groups, with their targets and four references fixed. Selection is independent of prediction outcomes but inherits the reference-first policy filters and possible MDT-training overlap. The seed-999 models remain fixed.

Both arms share context, target identity, reference solver and actual-forward MDT policies, and source-defined features. Columns are the effective-input union of at most three compared routers, at most 15 fields per case; selected oracle EV/action-conditioned quantities remain. Target solver/MDT

policies, route, logits and contributions are hidden. For the anchor and each shadow, treatment adds their first divergent greedy router, or the deepest common router marked non-divergent. It displays node reach, effective inputs, branch probabilities, signed additive raw-score differences and branch-conditional descendant action mixtures. Control keeps all shared evidence but omits these contrasts. All 80 reference forwards reconstruct routing, legally masked mixtures and score differences; maximum router discrepancy is 6.067×10^{-7}. Actual effective channels include fewer-than-five routers. Target-gold mutation, seven CPU regression tests and pre-call hashes check pairing, algebra, validation and accounting.

Execution and results. DeepSeek-V4 Pro uses the official wrapper, non-thinking, temperature zero, JSON and a 512-token cap: 40 single attempts, balanced ten per pair order, no retries. All distributions are valid; legal-uniform fallback is unused. Primary error is summed L_1; the 20,000-draw paired-board bootstrap uses seed 20260920. Control minus treatment is 0.023676, 95% interval

Table 42: **Compact contrasts with reference identities fixed.** All 20 targets are retained; lower error is better. Numerical baselines use the displayed labels.

Input	Summed L_1	Action-averaged L_1
Same labels and features	0.354859	0.118286
Add short reference contrasts	0.331184	0.110395
Four-reference mean	0.405485	0.135162
Legal-uniform	0.479779	0.159926

$[-0.057969, 0.120619]$, a 6.67% relative point reduction. Five improve, three worsen and twelve tie at 10^{-12}. Input totals are 29,910 versus 62,848 tokens ($2.10\times$). The interval is inconclusive and length is unmatched. The full-path control and the 80-target repeated-evidence comparison test different displays and populations.

G.4 Compact Contrasts with a Repeated-Evidence Control

To distinguish contrast information from additional prompt length, the remaining 80 reference-first targets (eight per scenario/seat subset) compare compact contrasts with repeated evidence, separately from the 20-target pilot. IDs, references, prompts, labels, settings and analysis were fixed before requests. Reference-first solver-label filters and possible training overlap remain.

Matched inputs. A supplies context, target features and four original reference solver/MDT policies with named features. B adds the pilot's three compact reference contrasts; C cycles existing reference facts to approximately match B's token length. No arm receives target policy, MDT mixture, route, logits or contributions. Only references enter model inference, and target-gold mutation leaves prompts unchanged. Oracle EV quantities and historical proxies retain source-defined units.

All 320 reference forwards reconstruct unmasked latent routing, legal leaf mixtures and additive scores; maximum discrepancies are 3.84×10^{-7}, 2.50×10^{-7} and 1.53×10^{-5}, respectively. Seventeen offline tests cover pairing, inference boundaries, validation and accounting. Official DeepSeek-V4 Pro uses temperature zero, thinking disabled, JSON and 512 output tokens. The 240 requests are single-attempt, with randomized cases and adjacent arms; each of six orders occurs 13 or 14 times. There are no retries, replacements, interim effect checks or prompt revisions. The model name is a provider alias.

Analysis. Primary A-minus-B and secondary C-minus-B compare summed L_1, positive favoring B. The strict three-slot parser requires finite legal probabilities summing within 10^{-6} of one and rejects duplicate keys. Invalid or missing outputs use legal-uniform fallback. All 80 targets remain. The 50,000-draw bootstrap (seed 20260921) resamples 79 exact sorted-card groups, preserving all cases and equal case weighting. Reported 95% intervals are descriptive; benefit beyond both controls requires both Bonferroni 97.5% lower bounds above zero. The analysis uses only this experiment's targets.

A minus B has 80 valid pairs; their sensitivity mean is -0.014360. C minus B has 80 valid pairs; their sensitivity mean is -0.028526.

Table 43: **Complete fixed-population confirmation.** All 80 targets remain in each arm. Lower error is better. Action-averaged error divides summed error by the three output slots.

Input	Valid	Summed L_1	Action-averaged L_1
Same labels and features	80/80	0.299886	0.099962
Add verified contrasts	80/80	0.314246	0.104749
Add redundant facts	80/80	0.285720	0.095240
Four-reference mean	—	0.365390	0.121797
Legal-uniform	—	0.487062	0.162354

Table 44: **Paired confirmation contrasts.** Positive favors verified contrasts. Wins/ties/losses use tolerance 10^{-12}.

Comparison	Mean	95% interval	Adjusted 97.5%	W/T/L
A minus B	−0.01436	[−0.06242, 0.02663]	[−0.07048, 0.03245]	17/49/14
C minus B	−0.02853	[−0.07304, 0.00653]	[−0.08057, 0.01044]	15/43/22

Prompt lengths. Provider-reported input totals for A, B and C are 116,252, 248,899 and 248,717 tokens. The pinned local tokenizer limits B/C length mismatch to 2% per case. Actual provider usage is available for 80 B/C pairs. Mean and maximum relative mismatches are 0.361% and 1.217%, respectively. Token matching does not remove semantic or presentation differences. Both

Table 45: **All ten predeclared subgroups.** Eight targets per row; summed errors. Subgroup differences are descriptive, not separately powered confirmation tests.

Scenario/seat	A	B	C	A–B	C–B
BTN_vs_BB_3B_IP	0.5024	0.4038	0.4426	0.0986	0.0388
BTN_vs_BB_3B_OOP	0.2169	0.2185	0.2012	−0.0015	−0.0173
BTN_vs_BB_SRP_IP	0.2436	0.2330	0.2414	0.0106	0.0084
BTN_vs_BB_SRP_OOP	0.2078	0.3728	0.2184	−0.1650	−0.1544
BTN_vs_SB_3B_IP	0.2777	0.2731	0.2543	0.0047	−0.0187
BTN_vs_SB_3B_OOP	0.3603	0.3728	0.3603	−0.0125	−0.0125
SB_vs_BB_3B_IP	0.2501	0.3691	0.3151	−0.1190	−0.0540
SB_vs_BB_3B_OOP	0.4544	0.4991	0.4244	−0.0447	−0.0748
SB_vs_BB_SRP_IP	0.2909	0.2004	0.2264	0.0905	0.0259
SB_vs_BB_SRP_OOP	0.1946	0.1999	0.1731	−0.0053	−0.0268

adjusted intervals include zero; the criterion for benefit beyond both controls is unmet. The evidence establishes neither equivalence nor universal harm. This experiment measures compact-path effects on a filtered population; it neither estimates the eight-configuration aggregate nor varies reference selection.

G.5 Target-Connected Paths on Reassigned Queries

Population. This experiment tests paths that connect the query to its references, rather than contrasts among references alone. From each reference-first scenario/seat subset, a fixed hash chooses six uniform-reference cases (Appendix H.2), totaling 60. The first reference becomes the query; the other three and the former query supply its four references. Reassignment precedes inspection of the query's policy, mixing, model agreement or error. The hands appeared in the reference-selection experiment, and the public nodes inherit its filtering and possible fitting overlap. The ten models remain fixed.

Three inputs. A supplies context, legal slots, query/reference identities, four reference solver/MDT policies and identical feature tables. Columns union the effective router inputs (29–49 per checkpoint); source names, formulas and dependency codes distinguish oracle inputs, historical proxies and provider scores. Missing values are null; display uses six significant digits. B adds target/reference soft-path relations. C repeats existing reference facts to match B within 2% local tokens, with

unchanged instructions and payload keys. This controls length/redundancy, not every linguistic difference.

Each path display is the shortest descending-reach prefix retaining 95% mass, ties by leaf ID, with omitted mass and visited-router input IDs. Four ternary digits identify latent branches. Overlap (Equation 4) uses all 81 paths, including the tail. No target solver policy, target MDT prediction, leaf strategy or suggested target distribution is transmitted. The added overlap is a computed relation, so performance evaluates use of this numerical interface.

Verification, execution and analysis. Independent scalar traversal checks all 81 weights for 300 hands, retained mass, overlap and row identity. A batch-of-five forward agrees within 2×10^{-6}. Gold mutation leaves inputs unchanged; removing treatment fields yields identical common evidence. Protocol, inputs and analysis are fixed before requests. Official DeepSeek-V4 Pro uses temperature zero, thinking disabled, JSON, 128 output tokens and 180-second timeout. The 180 serial requests receive one attempt each; six arm orders occur ten times, shuffled with cases using seed 202609211. All responses are valid; the model is an unpinned alias.

All 60 cases enter summed L_1. The strict three-slot parser requires finite, nonnegative legal probabilities summing within 10^{-6} and no duplicate keys; invalid/missing/truncated outputs use legal-uniform fallback. A-minus-B and C-minus-B use 50,000 paired resamples of 60 exact-board groups, seed 202609211, with descriptive 95% and Bonferroni 97.5% intervals. Independent parsing, scoring and bootstrap reproduce both contrasts.

Table 46: **Matched soft-path results.** All 60 predictions are valid in each condition. No outputs are dropped or retried.

Evidence	Summed L_1	Input tokens	Output tokens
A: Labels and features	0.694191	178,763	1,924
B: Add soft paths	**0.623131**	238,957	1,934
C: Add repeated facts	0.667012	238,960	1,927

Table 47: **Paired improvements.** Positive favors B. W/T/L counts cases with smaller/equal/larger error after adding soft paths.

Comparison	Mean	95% interval	Adjusted 97.5%	W/T/L
A minus B	0.07106	$[-0.04865, 0.20021]$	$[-0.06590, 0.22001]$	22/26/12
C minus B	0.04388	$[-0.03210, 0.13426]$	$[-0.04101, 0.14793]$	16/34/10

The corresponding relative mean reductions are 10.2% and 6.6%. Both adjusted intervals include zero, so the predeclared criterion for incremental benefit beyond both controls is not met. The intervals establish neither equivalence nor a stable benefit. The analysis is restricted to these queries.

Sensitivity and deterministic diagnostics. Using the ten scenario/seat groups as descriptive bootstrap clusters also gives intervals spanning zero: 95% intervals are $[-0.04233, 0.18058]$ for A-minus-B and $[-0.03781, 0.12674]$ for C-minus-B. Leave-one-subset-out mean differences remain positive, ranging from 0.03676 to 0.11314 and from 0.01913 to 0.06867, respectively. Since every output is valid, valid-pair means equal the primary means. Provider token counts give a maximum B/C mismatch of 0.417% per case.

Fixed reference-mean, inverse-feature-distance and overlap-weighted interpolation give summed errors 0.672409, 0.652609 and 0.536442. Feature distance uses finite common features, scales each by its range across the query and four references, and averages absolute differences. Inverse-distance weights are normalized; exact-zero ties share mass equally, and no varying features gives the reference mean. Overlap interpolation normalizes the four κ values, using equal weights if all are zero. No powers, temperatures, feature sets or neighbor counts are fitted to these labels. The complete MDT's offline error is 0.358905; this prediction is withheld from every prompt. These diagnostics characterize the numerical information and are not additional primary tests.

Table 48: **All ten subsets.** Six queries per row. Differences across subgroups are descriptive.

Scenario/seat	A: Labels/features	B: Soft paths	C: Repetition
BTN_vs_BB_3B_IP	0.6711	0.4778	0.5445
BTN_vs_BB_3B_OOP	0.6380	0.4786	0.2993
BTN_vs_BB_SRP_IP	0.4955	0.4622	0.4855
BTN_vs_BB_SRP_OOP	0.6830	0.6771	0.5608
BTN_vs_SB_3B_IP	0.3376	0.3142	0.4042
BTN_vs_SB_3B_OOP	0.3516	0.6592	0.6249
SB_vs_BB_3B_IP	1.0228	0.6431	0.8598
SB_vs_BB_3B_OOP	1.3321	1.3424	1.3144
SB_vs_BB_SRP_IP	0.5837	0.3170	0.5837
SB_vs_BB_SRP_OOP	0.8265	0.8597	0.9930

The 80-target reference-only contrast test and this target-connected interface use different targets and inputs. Their absolute errors are not a paired interface comparison; neither test meets its joint benefit criterion.

G.6 Target-Connected Paths on Independently Sampled Nodes

Population and inputs. To test target-connected paths without reference-first node selection, we choose four of the ten sampled nodes per scenario/seat subset by the fixed node-hash order (Appendix C.3). The first of each node's 16 uniform positive-reach hands is the query; the next four are references. All 40 exact boards differ from the reference-first communication boards, with no policy, mixing, distance, agreement or error filter. Training overlap remains possible; the communication models remain fixed.

A supplies context, identities, legal masks, reference solver/MDT policies and source-defined features from the checkpoint's effective-input union. The prompt uses abstract check/call, fold and bet/raise slots without invented aggressive sizes; recorded history is unchanged. B adds query/reference 95%-mass path prefixes, omitted mass, router input IDs and all-81-path overlaps; C repeats existing reference facts to match B. Instructions match Appendix G.5. Target solver/MDT policies, leaf strategies and target recommendations are absent; oracle features remain equally available.

With model parameters and inputs fixed, scalar reconstruction checks all paths for 200 query/reference instances. Changing withheld target labels leaves prompts unchanged, and common evidence is identical across conditions. Protocol, 120 prompts, parser and analysis are frozen before responses. Official DeepSeek-V4 Pro uses temperature zero, thinking disabled, JSON and 128 output tokens, with case/arm order seed 202609213. All 120 single attempts return valid strict three-key, normalized legal distributions; there are no retries. Legal-uniform fallback is unused. Maximum B/C provider-token mismatch is 0.50%; the returned alias is `deepseek-v4-pro`.

Table 49: The 40-node communication test with independently sampled nodes, separate from the 60-query reassigned-target population. All outputs are valid. Lower summed action L_1 is better.

Evidence	Summed L_1	Input tokens	Output tokens
A: Labels and features	.648810	119,398	1,287
B: Add target-connected paths	**.561679**	154,202	1,301
C: Add repeated facts	.728171	154,169	1,294

Table 50: Predeclared paired contrasts. Positive values favor paths. W/T/L counts individual queries with lower/equal/higher error in B.

Contrast	Mean improvement	Adjusted 97.5% interval	W/T/L
A minus B	.087131	[-.073758,.282226]	15/15/10
C minus B	.166492	[.021944,.353104]	14/18/8

Uncertainty and baseline comparison. The two comparisons use 50,000 resamples of the 40 distinct board/query units, seed 202609213, with Bonferroni-adjusted 97.5% intervals. Paths reduce mean error by 13.4% versus common evidence and 22.9% versus repetition. The latter comparison supports improvement beyond this matched-length control; the former interval includes zero. The predeclared requirement that both intervals favor paths is therefore not met. This population is analyzed separately from the 60 reassigned queries, with its sample size fixed before inference.

On these same queries and four reference labels, uniform averaging gives summed $L_1 = .747516$, inverse-feature-distance interpolation .601183, and overlap interpolation .409939. The private full-MDT prediction, not supplied to the LLM, gives .225126. The overlap baseline remains stronger than the path-conditioned LLM, distinguishing information present in routing from its effective use in a textual interface. The reference count here is four, versus fifteen in Appendix H.4; their errors are not pooled or treated as directly comparable estimates.

Table 51: All ten scenario/seat subgroups in the independently sampled-node test; four fixed queries per row. Subgroup means are descriptive.

Scenario/seat	A: Labels/features	B: Paths	C: Repetition
BTN_vs_BB_3B_IP	0.8904	0.9654	1.0377
BTN_vs_BB_3B_OOP	0.3370	0.3620	0.6111
BTN_vs_BB_SRP_IP	0.7126	0.5652	0.5601
BTN_vs_BB_SRP_OOP	1.0540	0.9486	0.9736
BTN_vs_SB_3B_IP	0.7522	0.8522	0.9022
BTN_vs_SB_3B_OOP	0.6294	0.3977	0.6388
SB_vs_BB_3B_IP	0.5364	0.5610	0.5474
SB_vs_BB_3B_OOP	0.6508	0.2747	0.6508
SB_vs_BB_SRP_IP	0.7604	0.5300	0.7056
SB_vs_BB_SRP_OOP	0.1650	0.1600	0.6545

This target-first sample removes reference-first public-node selection while using the same communication models and licensed input data. Its unmet joint criterion and numerical baselines concern this fixed interface, without establishing training holdout or semantic comprehension.

G.7 Compact Paths and Reference Assignment on 300 Nodes

Population. This comparison tests whether a compact path display helps prediction and whether assigning paths to the correct references matters. Thirty nodes per scenario/seat subset are selected from the preprocessed flop data, one distinct exact board each, excluding the 198 reference-first and fidelity-evaluation boards. Eligibility requires two legal actions and sixteen positive-reach hands. Seed-202609216 SHA-256 orders subset, board, history, seat and node identity. Sixteen hands are sampled uniformly by identity-seeded randomness; first is query, next four references. All 300 identities are fixed before solver-label extraction, with no outcome filters or later sample expansion. These boards may overlap fitting of the unchanged retained MDTs.

Matched evidence and assignment control. A supplies shared context, masks, hand identities, source-defined features and reference solver/MDT policies. B adds target/reference 95%-mass path prefixes, omitted mass, router input IDs and full-81-path overlaps. C preserves target routing but deranges the four reference routing rows, recomputing overlaps while keeping labels/features fixed. Identity-seeded derangements have no fixed point; B/C share byte counts and values, differing only in assignment. A/B are not length matched. The compact display shares definitions and uses column-indexed rows, reference IDs and four-digit ternary paths. All arms request mixtures and distinguish branches from actions. Target solver/MDT policies, leaf policies and computed target recommendations are absent. Mutation tests check private target-policy independence. Scalar reconstruction of every query/reference path has maximum error below 7.1×10^{-8}.

Execution and analysis. Prompts and analysis rules are fixed before requests. Six within-node orders occur fifty times each. DeepSeek-V4 Pro uses temperature zero, thinking disabled, JSON and 128 output tokens: 900 single attempts, no retries. The provider alias is unpinned. Invalid/truncated/missing outputs receive legal-uniform fallback. A-minus-B and C-minus-B use 50,000 exact-board

draws, seed 202609216, with Bonferroni 97.5% intervals. The joint criterion requires both lower bounds above zero and primary mean gain at least 0.05. Planning used paired SD 0.513 from the 40-query test in Appendix G.6, approximately 80% power for effect 0.091 at two-sided $\alpha = .025$; this gives no equivalence guarantee. Only the specified nodes enter this analysis.

All 900 requests finish; A/B have 299 valid outputs each and C has 300. Independent parsing, scalar scoring and batched bootstrap reproduce every error and both intervals. Provider B/C input-token counts match exactly within every pair.

Table 52: Compact-path communication: all 300 fixed board/query units.

Evidence	Summed L_1	Valid	Input tokens	Output tokens
A: Common labels/features	.724758	299	791,316	9,757
B: Faithful compact paths	.684121	299	1,042,439	9,744
C: Shuffled reference paths	.867118	300	1,042,439	9,754

Table 53: Prespecified contrasts; positive improvement favors faithful paths.

Contrast	Mean improvement	Adjusted 97.5% interval	W/T/L
A minus B	.040637	[-.039180,.121256]	102/116/82
C minus B	.182997	[.099492,.268374]	94/148/58

The point reduction versus A is 5.6%; its interval spans zero and its mean is below the predeclared 0.05 threshold, so the joint criterion is not met. Correct assignment improves over shuffling. Descriptive feature interpolation obtains $.630276$, routing-overlap interpolation $.499352$, and private full-MDT prediction $.303343$. The latter is withheld from every prompt. Routing-overlap interpolation has lower error than the path-conditioned LLM in this comparison.

Table 54: All scenario/seat means; thirty fixed targets per row. Descriptive subgroups, not selected success tests.

Scenario/seat	A: Common	B: Faithful	C: Shuffled
BTN_vs_BB_3B_IP	.8490	.9094	1.0123
BTN_vs_BB_3B_OOP	.5701	.6751	.9667
BTN_vs_BB_SRP_IP	.7979	.6344	.8454
BTN_vs_BB_SRP_OOP	.5809	.6665	.7056
BTN_vs_SB_3B_IP	.9747	.7358	1.0078
BTN_vs_SB_3B_OOP	.7848	.6522	.7258
SB_vs_BB_3B_IP	.8683	.8349	.9373
SB_vs_BB_3B_OOP	.6311	.4346	.8692
SB_vs_BB_SRP_IP	.5170	.6703	.8487
SB_vs_BB_SRP_OOP	.6738	.6279	.7524

Faithful paths improve over A in six subsets and worsen in four; the primary estimand remains the full 300-query mean. The sample is separate from earlier pilots and does not establish MDT training holdout.

H Reference Selection and Numerical Evidence Use

H.1 Target-First Reference Construction

Fixed targets, full candidate pools. To examine reference construction without SCCS-conditioned target eligibility, we reuse all 100 public nodes from Appendix C.3. Each target is the first hand in that node's previously uniform 16-hand sample; its next hand is the anchor. The target and anchor are fixed by identities, without solver-policy filtering. All other positive-reach hands at the node are candidates, totaling 39,557 pool rows including targets. No target, anchor or node is replaced. We compare four-reference sets: the next four uniformly sampled hands, the four feature-nearest hands to the target, and the anchor plus three shadows selected by SCCS's reference-construction component.

Reference-construction rule. Using the retained MDT, the component requires anchor–shadow policy $L_1 \geq 0.05$, prioritizes deepest first divergence of greedy routes, breaks ties by raw-feature distance, filters non-diverging candidates with the same most-probable-leaf action as the anchor's dominant policy action, and requires mutual shadow-policy $L_1 \geq 0.03$. The target is excluded from all candidate comparisons. Replay against the fixed forward outputs reproduces the shadow identities on all 100 nodes. This target-first test applies the reference-construction rule to positive-reach candidates with an independently fixed target. The reference-first protocol instead searches stored node rows and filters an evaluation target using solver labels (Appendix E.2). The anchor is independent of the target, and SCCS has no target-proximity term.

Predictors and fallbacks. Feature distance averages absolute differences over input columns observed throughout the full node pool, after pool-local min–max scaling and removal of constants. Nearest-reference ties are broken by cached row identity. For each set, use arithmetic label mean, inverse feature distance, or normalized 81-path overlap. Exact feature-distance ties at most 10^{-12} share all weight; zero overlap uses uniform weights. All selected IDs and weights are frozen before the scoring invocation opens solver labels. The target policy is not an input; solver-derived features and frozen model predictions are. Selection succeeds on 97 nodes. On three nodes with fewer than three valid shadows, use the already fixed uniform set and retain the target in every primary comparison.

Table 55: All 100 fixed targets, including fallback cases. Summed action L_1, lower is better.

Reference selection	Label mean	Inverse feature distance	Routing overlap
Uniform	0.711503	0.635464	0.471700
Target feature-nearest	0.160469	0.040372	0.130830
SCCS component	0.802173	0.778189	0.713448

Uncertainty and complete reporting. The predeclared primary contrast is SCCS minus uniform under inverse-feature weighting. Its mean is 0.142725, with 95% paired exact-board bootstrap interval $[0.021840, 0.266743]$. Positive means higher SCCS error. Across all six contrasts, we additionally report Bonferroni 99.1667% intervals from the same 50,000 draws, seed 202609215. The primary interval and the family-adjusted interval answer different multiplicity conventions; the latter spans zero. All targets and all nine predictor/selector combinations are retained. Conditioning descriptively

Table 56: SCCS minus comparator: six contrasts on all 100 targets.

Predictor	Comparator	Difference	Adjusted interval
Label mean	Uniform	0.090670	$[-0.052138, 0.232104]$
Label mean	Nearest	0.641705	$[0.466456, 0.823923]$
Inverse feature	Uniform	0.142725	$[-0.020812, 0.310152]$
Inverse feature	Nearest	0.737817	$[0.572910, 0.909618]$
Routing overlap	Uniform	0.241748	$[0.059715, 0.427380]$
Routing overlap	Nearest	0.582618	$[0.421292, 0.755112]$

on the 97 successful selections gives label-mean, inverse-feature and routing errors of 0.826947, 0.802230, 0.735483 for SCCS; 0.733472, 0.655092, 0.486259 for uniform; and 0.165421, 0.041619, 0.134866 for nearest. The primary result includes all targets. Supplementary records contain per-target errors, all ten subgroup means and reference weights. Independent forward reconstruction of all 1,600 model predictions and path weights agrees within 2×10^{-6}. These are numerical predictions from fixed models.

Nearest-reference characteristics. After scoring, we inspect the fixed nearest references without re-selection. None of the 100 targets has a zero-distance reference or an identical full feature/mask row. The closest reference shares the target's two card ranks in 89 cases. The low nearest-reference error therefore concerns a dense same-node pool of solver-derived summaries with many rank-related hands, not a demonstration of semantic reasoning or performance with sparse references on unseen public contexts.

Interpretation. The conditional advantage of SCCS references in Appendix H.2 does not carry over to this target-first numerical task. SCCS constructs contrasts between an anchor and shadows with different policies; retrieving the closest labels for an independently chosen target is a different objective. This comparison measures numerical label transfer on the stated population; it does not establish training holdout or LLM comprehension.

H.2 Reference Selection on Policy-Filtered Targets

This comparison tests reference selection while keeping the evidence format fixed. Each SCCS four-reference set is compared with four uniformly sampled distinct hands from the same public decision node, excluding the target and requiring positive reach. Uniform sampling does not inspect policy labels and may overlap the SCCS references. The 100 reference-first targets and MDT checkpoints remain fixed. Because these targets were filtered relative to SCCS references, the result is conditional on that selection. Both arms receive identical public context, legal actions, target summaries and definitions, and exactly four reference solver/MDT policies with the same input columns. Neither receives paths, contrasts or the complete target MDT policy. Input columns are fixed to those of Appendix F.5; the comparison therefore also conditions on that display's inputs. Only reference identities and their values differ. The target solver policy is hidden and the models remain fixed.

Fixed execution and analysis. All 200 prompts were frozen before any comparison outcome, using official DeepSeek-V4 Pro, temperature zero, thinking disabled, JSON mode and a 128-token cap. There is one attempt per target and arm, no retry or replacement. Per-case uniform sampling uses the integer SHA-256 digest of a fixed seed string and case ID; case order and 50/50 within-case arm order use seed 20260924. The model is a provider alias, not a pinned weight release. Invalid outputs use the same strict parser and legal-uniform fallback as Appendix F.5, retaining all 100 cases. The sole primary difference is selected minus uniform summed action L_1; negative favors SCCS selection. We use 50,000 paired case-weighted bootstrap draws over 98 exact board groups with seed 20260924 and a 95% interval. Independent parsing, scoring and interval recomputation agree. 0 outputs reached the cap. The paired mean difference is -0.318893 with 95% interval $[-0.439347, -0.198725]$.

Table 57: **Reference selection on policy-filtered targets.** All targets retained; lower is better.

References	Valid	Summed L_1	Label-mean L_1	Input tokens
SCCS-selected	100/100	0.365335	0.373409	729,325
Uniform same-node	100/100	0.684228	0.564918	729,455

Label-mean is the arithmetic mean of the four supplied solver distributions, computed without an LLM.

Table 58: **All ten reference-selection subgroups.** Ten cases per row; descriptive summed L_1.

Scenario/seat	SCCS-selected	Uniform
BTN_vs_BB_3B_IP	0.483358	0.574371
BTN_vs_BB_3B_OOP	0.423691	0.875971
BTN_vs_BB_SRP_IP	0.279353	0.371382
BTN_vs_BB_SRP_OOP	0.157913	0.798382
BTN_vs_SB_3B_IP	0.286588	0.625746
BTN_vs_SB_3B_OOP	0.415667	0.621666
SB_vs_BB_3B_IP	0.727141	0.843126
SB_vs_BB_3B_OOP	0.440573	0.727646
SB_vs_BB_SRP_IP	0.234817	0.708189
SB_vs_BB_SRP_OOP	0.204253	0.695804

Interpretation. The interval supports an advantage for SCCS-selected references under this fixed context and selected population. Both arms omit paths, so the comparison concerns reference selection alone. The target-selection filter, original feature set, single provider alias and one sampled

uniform set per target limit generalization. The path-contrast confirmation remains separate and is not pooled with this comparison.

H.3 Input-Only Label Interpolation

We test how accurately the supplied labels and numerical inputs can predict the target without an LLM. Three deterministic rules are fixed before inspecting the outcomes in Appendix H.2, with no fitted model, tuned bandwidth or target answer used for prediction. This descriptive comparison uses the same policy-filtered population and does not provide an independent generalization test. For a feature observed in the target and all four references, divide each target–reference absolute difference by that feature's range over those five hands. Ignore missing and constant columns, then average the scaled differences to obtain distance d_j to reference j. **Label mean** averages the four distributions. **Feature nearest** averages labels at the smallest distance (numerical tie tolerance 10^{-12}). **Inverse distance** averages with weights proportional to $1/d_j$; exact zero-distance references instead receive all weight equally. With no comparable columns, both distance methods return the label mean. The scaling is input-local, not learned from target outcomes. All three use the same solver-derived features and oracle reference labels available to the LLM.

Complete results and uncertainty. We retain all 100 targets under both reference sets. Paired differences are hosted LLM minus baseline summed action L_1; positive favors the baseline. Descriptive 95% intervals use 50,000 case-weighted paired bootstrap draws over 98 exact board groups, seed 20260927. They are not multiplicity-adjusted confirmatory tests. We report every predefined baseline rather than selecting a winner after seeing the outcomes. The hindsight oracle chooses the closest supplied solver label using the hidden target policy: it is an unavailable lower bound for copying one reference, not an implementable predictor. The LLM-versus-mean interval spans

Table 59: **All input-only baselines.** Lower L_1 is better; positive LLM-minus-baseline favors the deterministic comparator.

References	Predictor	Summed L_1	Difference	95% interval
SCCS	Hosted LLM	0.365335	–	–
SCCS	Label mean	0.373409	-0.008073	$[-0.09245, 0.07550]$
SCCS	Feature nearest	0.328837	0.036498	$[-0.01430, 0.09095]$
SCCS	Inverse distance	0.269905	0.095430	$[0.03499, 0.15757]$
SCCS	Hindsight label oracle	0.141751	0.223585	$[0.15119, 0.30320]$
Uniform	Hosted LLM	0.684228	–	–
Uniform	Label mean	0.564918	0.119311	$[0.02247, 0.21547]$
Uniform	Feature nearest	0.714869	-0.030641	$[-0.11981, 0.05654]$
Uniform	Inverse distance	0.571870	0.112359	$[0.02412, 0.19970]$
Uniform	Hindsight label oracle	0.329698	0.354531	$[0.25670, 0.45184]$

Table 60: **Every subgroup and implementable baseline.** S/U denotes SCCS/uniform references; ten targets per row. Entries are summed L_1.

Group	Mean S	Nearest S	Inverse S	Mean U	Nearest U	Inverse U
BTN_vs_BB_3B_IP	0.3665	0.3687	0.3042	0.5474	0.5773	0.5454
BTN_vs_BB_3B_OOP	0.4806	0.4140	0.4178	0.4926	0.6283	0.5200
BTN_vs_BB_SRP_IP	0.3150	0.3008	0.1898	0.2678	0.5835	0.3043
BTN_vs_BB_SRP_OOP	0.2967	0.2355	0.1428	0.7957	0.8358	0.7627
BTN_vs_SB_3B_IP	0.3624	0.1887	0.1883	0.5944	0.7256	0.6371
BTN_vs_SB_3B_OOP	0.3326	0.4146	0.3376	0.5451	0.7701	0.5640
SB_vs_BB_3B_IP	0.3803	0.3673	0.3212	0.6465	0.7185	0.6651
SB_vs_BB_3B_OOP	0.4780	0.4596	0.4110	0.5686	0.9079	0.5708
SB_vs_BB_SRP_IP	0.4452	0.3480	0.2792	0.6078	0.6746	0.5593
SB_vs_BB_SRP_OOP	0.2769	0.1911	0.1072	0.5833	0.7270	0.5902

zero; inverse-distance interpolation has lower error with a favorable descriptive interval. This result concerns the fixed solver-assisted inputs and policy-filtered population, leaving LLM necessity and

path utility unestablished. Six implementation tests cover independence from withheld target labels, permutation, units, missing/constant inputs, ties and exact neighbors. A scalar implementation reproduces every baseline loss.

H.4 Routing Overlap and Assignment Controls

Question and sample. Does the learned routing relation identify useful same-context reference policies beyond raw feature distance or an arbitrary assignment of the same weights? We reuse the fixed 100-node, 1,600-hand population of Appendix C.3. Each hand is predicted from the other 15 hands at the same node. Neither the target label nor the target MDT policy is an input to any weighting rule. Nodes and hands are not accepted or rejected based on their labels. This tests numerical information in the retained routing representation, separately from SCCS reference retrieval and separately from LLM interpretation.

Four fixed rules. All methods interpolate the same 15 solver-labeled reference distributions. Uniform averaging assigns weight $1/15$. Feature distance uses all 156 cached summary columns, retaining those unmasked and finite for every hand at the node and dropping constant columns. Each retained column is scaled by its range across the 16 hands. We use mean absolute feature distance to the query and normalize inverse distances to sum to one; exact zero-distance ties share all weight equally. If no varying column remains, the rule returns uniform weights. Feature scaling uses no labels.

For routing interpolation, all 81 path weights are normalized and the query–reference overlap is computed using Equation 4. Reference weights are overlap divided by total overlap; zero total overlap falls back to a uniform mean. The assignment control permutes the same weights across the 15 labels using a fixed identity-seeded rule: SHA-256 of `route-permutation:202609213:node-id:query-index`. Weighting rules are fixed before scoring, with no tuned temperature, exponent, neighbor count, metric or input selection.

Table 61: All-query numerical comparison on 100 distinct exact boards. Per-method intervals are 95% board bootstrap; the two predeclared paired contrasts use Bonferroni-adjusted 97.5% intervals.

Rule	Summed L_1	95% interval
Reference mean	.7097	[.6442,.7746]
Inverse feature distance	.5650	[.5135,.6158]
Routing overlap	.3985	[.3634,.4334]
Permuted overlap	.7118	[.6459,.7767]
Paired contrast	**Improvement**	**Adjusted 97.5% interval**
Feature minus overlap	.1665	[.1240,.2109]
Permuted minus overlap	.3133	[.2526,.3753]

Analysis and interpretation. We first average summed action L_1 over the 16 queries per node, then average across 100 nodes. The bootstrap resamples these board-level means 50,000 times with seed 202609213, preserving shared reference sets within nodes. Overlap improves the node mean on 72/100 nodes relative to feature distance and on 96/100 relative to permuted overlap. The 29.5% reduction versus feature distance and the assignment control support an informative learned routing relation. The checkpoint's supervised training is part of this representation; this experiment does not show that paths contain information unavailable to a supervised predictor.

These results concern retained checkpoints that may have seen these boards during fitting. Appendix C.1 tests routing on held-out boards using separately trained models; Appendix G.1 tests a fixed LLM's use of path displays. The numerical comparison here measures information in routing, with neither SCCS reference selection nor LLM interpretation as its treatment. It does not establish an advantage over dense learned similarities.

H.5 Complete-Output Similarity Controls

Question and information budget. To test whether the routing information in Appendix H.4 is also captured by complete policy outputs, we compare similarities derived from the fixed models' full action distributions. This exploratory analysis uses the same 100 nodes, 1,600 queries and 15 references per query, with no additional fitting. For each global teacher, tree teacher and MDT, normalize the three action probabilities and compute summed L_1 distance between query and reference predictions. Normalize inverse distances into reference weights; distances at most 10^{-12} instead receive equal weight, with all other weights zero. Interpolate the same reference solver labels. Target solver labels are not used in the weights, but the full target model output is used. That output includes the action head and was not supplied to the LLM; these are numerical supervised comparators, not information-identical LLM arms or capacity-matched sparsity ablations.

Fixed rules and results. All three weighting rules and their 4,800 weight vectors are fixed before scoring. This analysis is exploratory: the routing results on the same population were already known when the controls were designed. Each query is excluded from its references, and scalar summation verifies weighted predictions. Each exact board, containing 16 queries, is one bootstrap unit. We use 50,000 paired draws with seed 202609214 and Bonferroni 98.3333% intervals for the three comparisons. Negative differences favor the output-similarity control. All three complete-output

Table 62: Complete-output similarities on the unchanged 1,600-query numerical task.

Weighting	Summed L_1	Minus routing	Adjusted interval
Routing overlap	0.398465	–	–
Global-teacher output	0.190968	-0.207497	$[-0.243252, -0.172827]$
Tree-teacher output	0.314717	-0.083748	$[-0.111660, -0.056304]$
MDT output	0.327351	-0.071114	$[-0.092961, -0.050253]$

controls obtain lower error than routing overlap. These controls use the complete action distributions; routing overlap uses the distribution over leaves before the action head. The comparison measures predictive information in these representations on the same sampled-node population, which may overlap model fitting.

H.6 Prompt-Only Decoding and Mixture Preservation

Purpose and available information. This comparison asks whether a deterministic predictor can use the displayed evidence more accurately than DeepSeek-V4 Pro in the compact-path test (Appendix G.7). The decoder receives exactly the JSON in the 300-node prompts: legal slots, target/reference summaries, four reference solver labels and displayed overlaps. It has no model checkpoint, undisplayed path probabilities, target labels, target model outputs or leaf action distributions. All 1,200 predictions are computed before scoring against target labels. The comparison is exploratory because the DeepSeek outcomes were known when it was designed.

Parameter-free decoders. The overlap decoder normalizes the four displayed nonnegative overlaps into weights, using uniform weights when their sum is zero, and takes the corresponding convex combination of legally normalized reference labels. A feature decoder computes raw L_1 distance over columns observed in all five hands. A second rescales each such column by its query/reference range and removes constants, addressing unit sensitivity without fitting a metric. Exact matches within 10^{-12} share all weight; otherwise weights are proportional to $(d + 10^{-6})^{-1}$. With no usable columns, use uniform weights. We also apply the identical overlap decoder to the misassigned C prompt. No coefficients or thresholds are fitted to target errors.

Intervals use 50,000 paired exact-board bootstrap draws with seed 202609217 and Bonferroni correction across the three declared follow-up contrasts. All targets remain. The scaled-feature comparison shows that arbitrary raw units alone do not explain this decoder's advantage. It does not compare against a learned or optimally tuned feature metric.

Interpretation and reproducibility. The prompt itself suffices for lower-error numerical reconstruction than DeepSeek produces in this test. GPT obtains lower error than the same decoder in

Table 63: All 300 prompt-only decodes and DeepSeek-V4 Pro outputs in the compact-path test. Near-pure means largest action probability at least .95; 133 target policies are not near-pure. The last column counts collapse on those 133 targets.

Consumer	Summed L_1	Correct top action	Near-pure	On mixed targets
DeepSeek, common	.724758	173	140	40
DeepSeek, faithful paths	.684121	178	105	22
DeepSeek, shuffled paths	.867118	149	105	18
Raw-feature decoder	.630276	197	44	4
Scaled-feature decoder	.625831	201	43	4
Faithful-overlap decoder	.499352	219	70	8
Shuffled-overlap decoder	.797592	171	58	6

Table 64: Exploratory paired differences relative to faithful-overlap decoding. Positive means higher comparator error.

Comparator	Difference	Adjusted 98.333% interval
DeepSeek, faithful paths	.184769	[.095805,.277510]
Scaled-feature decoder	.126479	[.086730,.169412]
Shuffled-overlap decoder	.298240	[.218675,.380218]

Appendix F.6, so the ordering depends on the predictor. The decoder uses only the displayed evidence, without access to private target labels or unexposed action-head outputs. The comparison leaves the cognitive or architectural cause unresolved: the LLM and the decoder are different algorithms, and neither an information-theoretic ceiling nor human comprehension is measured. Both faithful and shuffled paths reduce DeepSeek's near-pure predictions relative to common evidence, while shuffled paths worsen policy error. Mixture preservation and correct assignment are therefore separate diagnostics; neither substitutes for full-distribution scoring.

The supplementary decoder operates directly on the displayed prompt. Synthetic tests check exact matches, zero-overlap fallback, invariance to changes of input units, exclusion of private-policy fields and rejection of invalid overlaps; they require no licensed hands or target answers.

I Response Diagnostics and Sensitivity

I.1 Output Concentration and Prototype Alignment

To characterize how the input conditions affect predicted mixtures, we analyze output concentration and proximity to displayed leaf prototypes. This post-hoc analysis uses all 100 targets and all four conditions from Appendix F.5. Near-pure means maximum action probability at least 0.99. Entropy is $-\sum_a p_a \log p_a$, divided by the log of the number of legal actions. Behavioral statistics use valid outputs only, with denominators shown; primary L1 includes all targets and the predefined fallback.

Table 65: **Valid-output concentration.** Counts retain their valid-output denominators; these are descriptive statistics.

Arm	Valid	Near-pure	Correct argmax	Mean entropy	Mean max prob.
Direct	100/100	17/100	47/100	0.7616	0.6510
Summaries	99/100	84/99	45/99	0.1397	0.9465
Route	100/100	87/100	50/100	0.0225	0.9959
SCCS	97/100	15/97	74/97	0.6256	0.7193

Table 66: **Distances to different prediction objects.** Valid-output mean summed L1, not solver-error estimates; lower means closer.

Arm	Greedy leaf	Full MDT	Label mean	Closer to leaf than MDT
Direct	1.045913	0.501231	0.551795	8/100
Summaries	1.099625	1.029750	1.074957	40/99
Route	0.239482	0.908374	0.995275	88/100
SCCS	0.866924	0.353565	0.397542	14/97

Table 67: **All subgroups: near-pure/valid counts.** Ten targets per group.

Group	Direct	Summaries	Route	SCCS
BTN_vs_BB_3B_IP	1/10	8/9	6/10	1/10
BTN_vs_BB_3B_OOP	0/10	5/10	9/10	2/10
BTN_vs_BB_SRP_IP	1/10	10/10	10/10	1/10
BTN_vs_BB_SRP_OOP	2/10	9/10	9/10	1/10
BTN_vs_SB_3B_IP	1/10	10/10	10/10	2/10
BTN_vs_SB_3B_OOP	4/10	9/10	8/10	2/9
SB_vs_BB_3B_IP	0/10	8/10	10/10	2/10
SB_vs_BB_3B_OOP	4/10	10/10	8/10	3/10
SB_vs_BB_SRP_IP	2/10	10/10	7/10	1/10
SB_vs_BB_SRP_OOP	2/10	5/10	10/10	0/8

Summaries and Route frequently produce near-pure distributions. Route is closer to its displayed leaf than the full MDT in 88/100 cases, consistent with prototype alignment. This association does not identify a cause: Summaries also concentrates without a leaf, and neither arm isolates the features, raw terms or instructions. No output ends for length. No rationale was requested; numerical alignment does not expose internal reasoning. Appendix I.2 tests the leaf field directly.

I.2 Single-Leaf Display Intervention

To test whether the displayed leaf contributes to the association in Appendix I.1, we compare paired prompts with and without that field. Four reference-first target IDs are sampled per scenario/seat subset using seed 20260928 and identifiers only, giving 40 targets. Each receives two single-attempt predictions: the verified Greedy Route-only prompt with its leaf prototype, and the same prompt with only the `leaf_prototype` field removed. Route computations, target summaries, definitions and instructions are identical. Both prompts state that one leaf is not the complete mixture. Neither condition includes references, the complete target MDT policy or target solver answers. Both predictions are collected under this paired protocol. All 80 prompts and the analysis are frozen before calls. Official DeepSeek-V4 Pro uses temperature zero, thinking disabled, JSON mode and 128 output tokens, one attempt per case/arm with no retry or replacement. Case order is randomized and within-case arm order is balanced 20/20 with the fixed seed. The model is a provider alias, not a pinned weight revision. The single predeclared contrast is present minus absent summed action L_1, positive favoring removal. A paired, case-weighted exact-board bootstrap uses 50,000 draws over 40 groups with seed 20260928 and a 95% interval. All 40 cases remain under the existing strict parser and legal-uniform fallback. Concentration and subgroups are descriptive.

Table 68: **One-field intervention.** Both arms retain identical route evidence apart from the single leaf-distribution field.

Leaf prototype	Valid	Summed L_1	Near-pure/valid	Input tokens
Present	40/40	0.925763	35/40	207,044
Absent	39/40	1.107223	34/39	205,777

Present minus absent is -0.181460, with 95% interval $[-0.361823, 0.001656]$.

Table 69: **All intervention subgroups.** Four targets per row; descriptive summed L_1.

Group	Present	Absent
BTN_vs_BB_3B_IP	1.444902	0.691879
BTN_vs_BB_3B_OOP	0.884501	1.289992
BTN_vs_BB_SRP_IP	0.981959	0.983650
BTN_vs_BB_SRP_OOP	0.563386	1.184692
BTN_vs_SB_3B_IP	0.778273	1.343373
BTN_vs_SB_3B_OOP	1.165169	0.918052
SB_vs_BB_3B_IP	1.139712	1.474377
SB_vs_BB_3B_OOP	0.845762	1.000112
SB_vs_BB_SRP_IP	0.897411	1.131330
SB_vs_BB_SRP_OOP	0.556558	1.054771

Removal does not meet the improvement criterion: its point error is higher and the interval spans zero. Near-pure outputs persist in both arms. No output ends for length; one absent-leaf output receives fallback. Independent checks verify the one-field difference, errors and interval. This result leaves shared features, route terms and instructions unresolved; it neither establishes equivalence nor a benefit of retaining the leaf. The 40 filtered targets and one provider limit scope.

I.3 Clustering, Coverage and Subgroup Sensitivity

This post-hoc analysis tests sensitivity to the clustering unit while holding predictions and per-target errors fixed. We compute descriptive paired 95% percentile intervals using 50,000 resamples with seed 20260930, under exact-board, suit-equivalent-board and scenario/seat clustering. A suit-equivalent board is the lexicographically minimum sorted card sequence under all 24 suit relabelings. Scenario/seat has only ten clusters, so its interval provides a sensitivity check rather than reliable evidence of generalization to a new population.

Table 70: **All cluster sensitivity intervals.** Summed L1 difference uses the row order; lower error is better. Intervals are descriptive and not multiplicity-adjusted.

Comparison	Exact board	Suit-equivalent	Scenario/seat
Direct-SCCS	[0.0839, 0.3071]	[0.0853, 0.3073]	[0.0799, 0.3099]
Route-SCCS	[0.5657, 0.7443]	[0.5672, 0.7436]	[0.5455, 0.7630]
A-B	[-0.0624, 0.0266]	[-0.0620, 0.0267]	[-0.0644, 0.0320]
C-B	[-0.0728, 0.0065]	[-0.0725, 0.0064]	[-0.0641, 0.0014]
Selected-uniform	[-0.4391, -0.2003]	[-0.4384, -0.1986]	[-0.4337, -0.2076]
Present-absent	[-0.3632, -0.0002]	[-0.3622, 0.0008]	[-0.4055, 0.0878]

The 100-, 80- and 40-target studies have 98, 79 and 40 board groups, respectively. Suit canonicalization merges no additional groups here; the slight interval differences reflect finite resampling after group relabeling. The leaf comparison is sensitive to resampling near zero: its prespecified fixed-seed interval $[-0.361823, 0.001656]$ crosses zero, while this exact-board interval barely excludes zero. The wider scenario/seat interval also crosses zero, leaving the leaf effect inconclusive. Both contrast-control intervals remain inconclusive under every grouping.

Table 72: **Exact strategic evaluation in two complete local River endgames.** Values are milli-big-blinds per game (mbb/g), followed by exploitability as a percentage of the root pot. Hard–Soft is an observed deployment difference under different optimization settings, not an isolated sparsity effect. Deep CFR is a same-tree finite-training reference. Lower is better.

River state	Soft MDT	Hard Top-5 MDT	Hard–Soft	Deep CFR
Subgame 3 (5 bb pot)	6.691 (0.1338%)	17.173 (0.3435%)	10.482 (0.2096 pp)	43.075 (0.8615%)
Subgame 4 (37.5 bb pot)	55.013 (0.1467%)	133.046 (0.3548%)	78.033 (0.2081 pp)	133.495 (0.3560%)

Table 71: **Missing-output and subgroup sensitivity.** LOO is the range of means after omitting each scenario/seat subset. Invalid-loss bounds vary invalid errors over $[0, 2.00001]$ to allow source rounding; these are empirical bounds, not confidence intervals.

Comparison	Valid pairs / mean	LOO mean range	Invalid-loss bound
Direct-SCCS	97 / 0.1945	[0.1602, 0.2283]	[0.1523, 0.2123]
Route-SCCS	97 / 0.6680	[0.6289, 0.6838]	[0.6133, 0.6733]
A-B	80 / -0.0144	[-0.0269, 0.0024]	[-0.0144, -0.0144]
C-B	80 / -0.0285	[-0.0360, -0.0145]	[-0.0285, -0.0285]
Selected-uniform	100 / -0.3189	[-0.3442, -0.2832]	[-0.3189, -0.3189]
Present-absent	39 / -0.1724	[-0.2853, -0.1326]	[-0.2123, -0.1623]

Package and conditional selected-reference gains survive subgroup omission and fallback sensitivity. Original filtering, fitting overlap and provider/solver dependence remain. These overlapping studies are not independent replications. Primary adjusted intervals remain those in Appendices G.4 and F.5; the leaf intervention retains its original inconclusive interval.

J Strategic Value and Complete River Endgames

J.1 Local Mixture Error and Fixed-Opponent Value

L_1 is primary and local EV is auxiliary. The action-averaged L_1 directly measures fidelity to the solver's complete local mixture. Its unaveraged form is twice total variation, so probability-mass displacement changes the metric linearly. By contrast, the oracle-conditioned EV gap evaluates the predicted mixture using fixed solver continuation values. At an exact equilibrium information set, every supported solver action is value-maximizing; under that condition, the one-decision EV gap equals the predicted mixture's one-step regret relative to the best action. Our targets are approximate solver solutions, so we use the more limited term *local value gap*. It is not CFR counterfactual regret or a complete-policy best-response metric.

EV alone is insufficient for the articulation objective. Multiple supported equilibrium actions can be nearly indifferent against the fixed equilibrium opponent, so an almost-pure prediction can have low local EV loss while remaining far from the intended mixture. L_1 therefore determines checkpoint selection, while the EV term discourages locally costly probability shifts among predictions with similar distributional fidelity.

Corpus-scale fixed-opponent comparison. As a complementary NLH diagnostic, we evaluate Soft-sparse Tree and final Hard MDT against the fixed approximate-Nash opponent on the solver-labeled postflop distribution. Their value losses are 7.836 and 20.688 mbb per evaluated postflop hand, respectively, a descriptive difference of 12.852 mbb per evaluated postflop hand. This evaluation keeps the opponent and solver-derived summaries fixed. The checkpoint families differ in more than the Top-5 mask. The River study additionally matches architecture and evaluates exact best responses, but its different optimization schedules likewise prevent isolating a mask-only effect.

J.2 Complete-Tree Best Responses and Deployment Costs

Public states and action trees. The experiment starts from the released HUNL River states used for Subgames 3 and 4 by Brown & Sandholm (2019a). Subgame 3 uses board `4s 8h Tc 9h 2s`, a 5

bb root pot, and 197.5 bb remaining per player. Its two root bet sizes are $3.4/7.5$ bb and its two bet sizes after a check are $2/3.4$ bb. Subgame 4 uses board `Js Ks 5c Qs 7d`, a 37.5 bb root pot, and 181.25 bb remaining per player. Its corresponding sizes are $9.4/37.5$ bb at the root and $56.3/112.5$ bb after a check. We retain the released root ranges and build complete local trees through fold or showdown. These are our own two-size trees initialized from the published states, not the betting abstraction used in the original benchmark curves.

Local DCFR solves provide high-precision supervised targets. Their residual exact exploitabilities are 0.009034 mbb/g on Subgame 3 and 0.009907 mbb/g on Subgame 4. Both MDT variants use depth-4 ternary probabilistic routing and pure-action leaves, with one model per acting position shared across that position's decision states. The inputs exclude action Q/value features, target-policy probabilities, and exact private-card indicators. Soft can use all eligible summaries; Hard uses exactly five summaries at each router. After supervised fitting, the deployments are fine-tuned and selected using the exact same-tree strategic metric. The result is therefore an in-domain strategic-fit evaluation of the compressed policies.

Tree construction and strategic optimization. Tree construction specifies two dynamic bet and raise candidates per player, a maximum of two raises, and no automatic extra all-in. At subsequent states, legal bounds or merging can leave one aggressive size; sizes are sorted into small/large slots alongside fold, check and call. River fold and showdown children terminate traversal. Reproduction requires the resulting action trees as well as these construction settings.

For strategic fine-tuning, each outer step recomputes both information-set best responses on the complete tree. Holding those discrete responses fixed, Adam minimizes

$$J(\theta) = V_0(\mathrm{BR}_0, \pi_{\theta,1}) + V_1(\pi_{\theta,0}, \mathrm{BR}_1) + \lambda L_{1,\mathrm{target}}.$$

The first two terms equal NashConv at the current response anchor in the two-player zero-sum game. Values use big blinds; $L_{1,\mathrm{target}}$ averages absolute error over legal actions with equal-public-node weighting and root-range hand weights within a node. The selected final continuations take one optimizer step per refreshed response and clip gradient norm at 5. Their router supports and pure leaf assignments are fixed. Selection uses freshly recomputed exact exploitability, not the fixed-response surrogate.

Table 73: **Final continuation settings for selected River policies.** Budgets refer to the last continuation, not the full supervised fitting and earlier search. All selections use the evaluated tree, not an independent test tree.

State	Variant	Learning rate	λ	Rounds	Selected round
3	Soft	0.003	1.0	100	95
3	Hard	0.01	0.1	200	200
4	Soft	0.0003	1.0	500	469
4	Hard	0.001	1.0	400	390/399

The final Hard Subgame-4 deployment combines actor 0 from round 390 with actor 1 from round 399, selected by their respective opponent best-response terms. A full exact recomputation accepts this combination at 133.046 mbb/g, versus 133.064 for the best same-round pair. These terminal settings and selection provenance clarify the reported in-domain fit; they are not an equal-total-compute comparison with Deep CFR or a held-out estimate of strategic generalization. Soft and Hard also have different continuation budgets and hyperparameters. Their reported gap is an observed difference between selected deployments, not a mask-only causal effect under matched optimization.

For the matched learning-based reference, Deep CFR uses $10{,}000$ sampled game-tree traversals per player per iteration. We evaluate its linearly weighted empirical-average strategy, rather than fitting a separate final actor, with the same exact best-response computation. At iteration 800 it reaches 43.075 mbb/g on Subgame 3 and 133.495 mbb/g on Subgame 4. These checkpoints give scale on the same trees and metric; they are neither convergence limits nor a general comparison between MDT and Deep CFR outside these two in-domain endgames.

Table 74: **Additional River-endgame diagnostics.** L_1 is equally averaged over decision states. Fixed-target value loss keeps the local target opponent fixed; exact exploitability recomputes each player's best response. All value entries are mbb/g.

State	Policy	L_1	Fixed-target loss	Exact exploitability
Subgame 3	Soft MDT	0.01214	1.909	6.691
	Hard Top-5 MDT	0.03751	7.704	17.173
Subgame 4	Soft MDT	0.02746	25.917	55.013
	Hard Top-5 MDT	0.05964	54.951	133.046

K Cross-Game Validation in Liar's Dice

K.1 Private-Type Cross-Fitting and Fixed-Opponent Evaluation

Game and protocol. The game has two players, two three-faced dice per player, and bids up to quantity three. Because a player's two dice are unordered and repetitions are allowed, there are six private types: $11, 12, 13, 22, 23, 33$. A full-tree CFR+ solver produces the approximate-Nash policy. The MDT interface contains 17 public-history, posterior-belief, private-state, and equity summaries; per-action continuation Q-values, Q-gaps, and the target policy are excluded from both MDT inputs and model-facing prompts. Six cross-fitting folds each hold out one private type at every public context. The communication evaluation contains 31 held-out target information states and four matched prompts per target. For each state h, the communication scorer reports legal-action-averaged error $|\mathcal{A}(h)|^{-1} \sum_{a \in \mathcal{A}(h)} |\hat{p}_a - p_a|$, using either the solver or MDT policy as p. The denominator is the state's number of legal actions; it can vary across states. This differs from both summed L_1 and the original NLH fixed-three-slot average. Configuration entries average these errors over parsed targets; the Mean row weights configurations equally. No scores are rescaled here.

The 31-target evaluation is exploratory and was designed after inspecting pilot results. Its eligibility rules, prompts and reference/shadow ranking were fixed before collecting these responses. Target eligibility required its solver policy to be at least 0.10 total-variation distance from both the training reference and shadow. These distances gated selection but did not rank eligible pairs. Metrics use successfully parsed outputs, with no imputation for missing responses. Table 75 reports all configuration-level results.

Six configurations contain all 124 expected responses. Gemini-3.1 Pro high contains 117 of 124, and GPT-5.5 high contains 123 of 124. In the order Direct, Summaries, Greedy Route-only, and SCCS, the parse counts out of 31 are: Gemini-3.1 Flash $(24, 27, 31, 15)$; Gemini-3.1 Pro low $(31, 31, 31, 31)$; Gemini-3.1 Pro high $(29, 28, 31, 29)$; DeepSeek-V4 Flash $(30, 31, 31, 30)$; DeepSeek-V4 Pro $(31, 31, 30, 29)$; GPT-5.4 and GPT-5.5 low $(31, 31, 31, 31)$; and GPT-5.5 high $(30, 31, 31, 31)$. The unweighted Mean row therefore summarizes configuration-level, parse-conditional point estimates rather than eight complete replications. On paired parse-success targets, the point estimate favors SCCS over Direct and Direct+Summaries in every listed configuration, whereas the smaller SCCS–Greedy Route-only margins are heterogeneous and should be interpreted as a component diagnostic.

Complete fixed-opponent value evaluation. The six held-out folds provide exactly one MDT prediction for each of the game's 3,072 information sets. We assemble these predictions into a fixed behavioral policy and traverse the complete tree against the approximate-Nash solver, once in each player position. Relative to solver self-play, the MDT value losses are 0.080723 as player 0 and 0.035129 as player 1, for a seat average of 0.057926 (reported as 0.058) per game on the terminal payoff scale $[-1, +1]$. This measures the complete policy's value against a fixed reference opponent; it does not recompute an adaptive best response.

L Evaluation Coverage and Reproducibility

Measured outcomes. MDT is the learned policy representation; SCCS supplies same-context contrast sampling and evidence construction for its in-context evaluation. LLM parameters remain

Table 75: **Per-configuration Liar's Dice strategy prediction.** Model entries are mean ± standard error over parse-success cases. Asterisks mark incomplete provider runs; missing responses are not imputed.

Solver target

LLM configuration	Direct	+Summaries	Greedy Route-only	SCCS
Gemini-3.1 Flash	.141 ± .019	.165 ± .022	.107 ± .012	.096 ± .015
Gemini-3.1 Pro low	.171 ± .020	.149 ± .017	.132 ± .014	.123 ± .011
Gemini-3.1 Pro high*	.168 ± .019	.148 ± .014	.125 ± .011	.100 ± .011
DeepSeek-V4 Flash	.266 ± .015	.270 ± .014	.134 ± .012	.120 ± .012
DeepSeek-V4 Pro	.280 ± .012	.280 ± .012	.128 ± .014	.111 ± .013
GPT-5.4	.113 ± .013	.120 ± .012	.094 ± .012	.088 ± .012
GPT-5.5 low	.112 ± .014	.113 ± .017	.101 ± .013	.094 ± .011
GPT-5.5 high*	.137 ± .018	.124 ± .015	.105 ± .013	.104 ± .012
Mean	.174 ± .065	.171 ± .067	.116 ± .016	**.105 ± .012**

MDT target

LLM configuration	Direct	+Summaries	Greedy Route-only	SCCS
Gemini-3.1 Flash	.154 ± .023	.167 ± .026	.087 ± .010	.068 ± .009
Gemini-3.1 Pro low	.192 ± .026	.167 ± .023	.095 ± .013	.078 ± .009
Gemini-3.1 Pro high*	.184 ± .025	.148 ± .020	.090 ± .009	.080 ± .010
DeepSeek-V4 Flash	.271 ± .019	.273 ± .018	.089 ± .009	.072 ± .009
DeepSeek-V4 Pro	.285 ± .016	.285 ± .016	.089 ± .009	.088 ± .011
GPT-5.4	.124 ± .017	.126 ± .017	.072 ± .006	.077 ± .008
GPT-5.5 low	.131 ± .018	.104 ± .015	.068 ± .006	.083 ± .008
GPT-5.5 high*	.153 ± .024	.120 ± .016	.068 ± .007	.086 ± .008
Mean	.187 ± .061	.174 ± .069	.082 ± .011	**.079 ± .007**

fixed. Hidden-mixture prediction, computational reconstruction and strategic policy value are distinct outcomes; no direct semantic-comprehension or human-inspection measure is collected. Adding reference-path displays to shared solver labels, numerical summaries and reference MDT predictions has not met the adjusted incremental-benefit criteria on training-held-out boards. Nearest-reference retrieval is stronger than SCCS in the target-first package test.

Data and information access. The grouped study covers five scenarios, three seeds on one partition and a solver-derived input interface. Communication checkpoints trained without board grouping may overlap the evaluation boards; reference-first targets also use policy filters and the documented rendering conventions. The model's local input bound does not imply a globally five-input explanation. NLH summaries require solver access, so the present system is an offline analytical interface rather than a standalone live agent. Predictions use solver-supported response actions; robustness to arbitrary histories or action sizes is untested. SCCS may fail to find feasible shadows; the target-first study retains its prespecified uniform-reference fallbacks. Benefits outside the tested mixed-strategy games are unknown.

Reproducibility. Bulk redistribution of the commercial NLH corpus is restricted. The interface schema, implementation, configurations, prompts/evaluation code, aggregates and legally distributable examples can be released, but regenerating the 250-million-decision corpus requires a compatible licensed solver. Liar's Dice provides a releasable solver, data generator, targets, cross-fitted MDTs, prompts, outputs and scoring code. The corpus-scale tables identify unavailable dispersion definitions; the cross-model tables report the available execution settings and response coverage.