ENDOGENOUS CURRICULA FOR VERIFIABLE SELF-PLAY: PROJECTED SOFT-Q COVERAGE–INFORMATION CERTIFICATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-play agents must adapt executable-task curricula without extinguishing valid program paths. We model this as KL-regularized control on a syntax-constrained prefix tree. The same projected soft-Q iterate yields a completed-program floor and an information-value certificate. A realized-row bound sharpens support control; direct performance identities, quadratic residual penalties, and contrast-aware estimator errors certify information advantage over an IG-ablated comparator without the full-objective optimal-policy oracle. A coherent finite Bayesian specialization gives expected entropy budgets, while identifiable likelihoods and accumulated row floors yield anytime posterior and finite-time identification bounds. Fixed-horizon two-time-scale tracking and epochwise expansion delimit the dynamical scope. In an executable Boolean-DSL study, a matched pure-information curriculum improves target success and semantic coverage over its reference. Shared-belief audits show conditional gains after KL costs, but the original residual-composed information bound is nonpositive throughout the tested configurations. The direct refinements have not been evaluated on those records; we separate mathematical guarantees, observed learning gains, and unresolved finite-time efficacy.

1 INTRODUCTION

A self-play agent can generate executable tasks, receive grounded feedback, and learn from the resulting curriculum. The task distribution then depends on what the learner knows. Self-play and automatic curricula motivate this interaction (Sukhbaatar et al., 2018; Portelas et al., 2020; Chen et al., 2025), but syntactic support, successful execution, and informative supervision are distinct. A syntax mask does not prevent concentration within the legal language, and uncertainty rewards alone do not quantify the loss from approximate generator optimization.

We ask how one implemented generator can preserve valid program paths while retaining the information value of its curriculum. The model combines a full-support legal-token reference, bounded verifier/frontier/novelty and information-gain (IG) rewards, and a projected soft-Q table. The joint certificate in Theorem 4.12 uses the *same* iterate: projection supplies a completed-program floor regardless of convergence, and a Bellman residual controls information loss through terminal-law tracking. A realized-row refinement keeps actual action gaps rather than charging every token the projection width.

Direct certificates and learning. Theorem 4.14 bypasses the full-objective optimal-policy oracle in the information comparison. Exact objective identities, quadratic residual penalties, and contrast-aware surrogate errors quantify advantage over an IG-ablated comparator. A separate finite-hypothesis result turns accumulated row floors and likelihood separation into anytime posterior and finite-time identification bounds. These use standard control and martingale tools to connect implemented generation to learning; they do not prove that IG is necessary or uniformly faster than reference sampling.

Dynamics and evidence. We distinguish fixed-horizon two-time-scale tracking, epochwise expansion, and the discrete finite-Bayes specialization. An executable 16-hypothesis, 192-program

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.

study evaluates policies on shared beliefs for objective comparisons and on separate rollouts for learning curves. It shows conditional learning gains, but the original residual-to-TV substitution leaves no positive information bound in the tested configurations. The new direct bounds have not been evaluated on those records. This separates the demonstrated selection mechanism from the stronger finite-time guarantee still requiring evaluation.

2 Related Work

Regularized control. Soft Bellman and Gibbs results follow from regularized MDP theory (Geist et al., 2019; Haarnoja et al., 2018). Trust-region methods, PPO, and RLHF use related policy constraints or reference penalties (Schulman et al., 2015; 2017; Ouyang et al., 2022). We connect a projected table to support and information transfer, not a generic neural PPO update to tabular convergence.

Information and curricula. BALD expresses observation value through predictive entropies (Houlsby et al., 2011); dropout, ensembles, randomized priors, and epistemic networks offer uncertainty representations (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Osband et al., 2018; 2021). Our identity is exact for a finite module, with explicit surrogate errors rather than assumed neural calibration. Task selection is already endogenous in self-play and automatic curricula (Sukhbaatar et al., 2018; Portelas et al., 2020; Chen et al., 2025; Chae et al., 2025); we quantify its transfer through an implemented program generator.

Executable programs. Learned synthesis uses search, execution, or reusable programs (Balog et al., 2017; Devlin et al., 2017; Ellis et al., 2023; Chen et al., 2019); code-learning systems exploit tests and verifiable feedback (Le et al., 2022; Chen et al., 2023; Wen et al., 2025; Venkatkrishna et al., 2026). Our verifier-grounded subclass adds explicit prefix support and a projected decoder, the conditions enabling the joint certificate.

3 Formal Model

3.1 Program paths, observations, and state

Let Σ be a finite token alphabet and $\texttt{EOS} \notin \Sigma$ a termination token. A computable syntax mask assigns a nonempty finite legal-action set $\mathcal{A}(s) \subseteq \Sigma \cup \{\texttt{EOS}\}$ at every nonterminal legal prefix s. The legal tree is fixed across rounds. A horizon $L \geq 1$ allows at most L actions, including $\texttt{EOS}$. Write $\mathcal{S}_L$ for its prefixes and $\mathcal{X}_L$ for its nonterminal state–action pairs. A completed program u ends in $\texttt{EOS}$, has length $\ell(u)$ including that token, and belongs to the countable set $\mathcal{U}$.

Definition 3.1 (Reference policy and terminal laws)**.** The fixed reference satisfies $\pi_{\mathrm{ref}}(a \mid s) > 0$ on $\mathcal{A}(s)$ and sums to one there. Its probability of a completed path is $\pi_{\mathrm{ref}}(u) = \prod_{k<\ell(u)} \pi_{\mathrm{ref}}(a_k \mid s_k) > 0$. For fixed L, Ω_L contains all completed paths of length at most L and separately identified paths truncated at L. The probability law $\mu_\pi \in \Delta(\Omega_L)$ is induced by sampling every action from π. It is *not* conditioned on completion or verifier acceptance. Deterministic padding, when used, has probability one and preserves this law.

Definition 3.2 (Executor, verifier, and observation)**.** A sandboxed deterministic language $\mathcal{L}$ has a bounded executor $E(p, i; T) \in \Sigma^* \cup \{\bot\}$, with $\bot$ denoting failure to return an output within the budget. A computable verifier $V(x, y) \in \{0, 1\}$ checks the declared bounded specification. Acceptance asserts correctness only for that specification. A task map $x : \Omega_L \to \mathcal{X}$ includes a sentinel for failed or truncated outputs. The learner observes Y through a specified channel; a noisy observation of a sound verifier bit is distinguished from the verifier itself.

Definition 3.3 (Slow state and timing)**.** The round-t state z_t summarizes solver, buffer, staleness, and epistemic-module variables. For the ODE results only, $z_t \in \mathcal{Z} \subset \mathbb{R}^{d_z}$ is finite-dimensional and compact. The pre-generation state (q_t, z_t) and the frozen round objective are $\mathcal{F}_{t-1}$-measurable; the sampled output U_t, observation Y_t, and resulting updates belong to $\mathcal{F}_t$. Rounds start at $t = 1$. No sufficient-state or regularity claim for an arbitrary growing neural replay buffer is implicit in this definition.

3.2 INFORMATION SCORES AND KL CONTROL

Definition 3.4 (Finite epistemic module and implemented score)**.** For $M \in \{1, \dots, K\}$, let $w_z \in \Delta_K$ and let $p_{z,m}(\cdot \mid x)$ be predictors on a finite answer abstraction $\mathcal{Y}_x$. With $\bar{p}_z = \sum_m w_z(m) p_{z,m}$ and natural logarithms, define

$$I_z(x) := \widehat{\mathrm{IG}}_z(x) = H(\bar{p}_z(\cdot \mid x)) - \sum_m w_z(m) H(p_{z,m}(\cdot \mid x)).$$

Choose $I_{\max} \geq \sup_x \min\{\log K, \log |\mathcal{Y}_x|\}$. If $\widetilde{\mathrm{IG}}_z$ is the raw estimator, the *only* score entering control is $\widetilde{I}_z := \mathrm{clip}(\widetilde{\mathrm{IG}}_z, 0, I_{\max})$. Sentinel observations have a hypothesis-independent likelihood, so their scores are $I_z = \widetilde{I}_z = 0$. A neural uncertainty module can supply a surrogate, but the finite-module identity does not certify its calibration.

Definition 3.5 (Composite reward and finite-horizon objective)**.** Let $\lambda_I > 0$, $\lambda_N \geq 0$, and

$$R_z(u) = R_z^{\mathrm{other}}(u) + \lambda_I \widetilde{I}_z(x(u)), \qquad R_z^{\mathrm{other}} = r_F + \lambda_N r_N + r_V.$$

Frontier, clipped novelty, and verifier rewards are bounded, with

$$|R_z(u)| \leq R_{\max} < \infty. \tag{1}$$

We use terminal rewards unless specified otherwise: $r_z(s_h, a_h) = 0$ until termination, when the bounded terminal payoff is paid. With $0 < \gamma < 1$, stopping time $\tau_L \leq L$, and $\Gamma_L(u) := \gamma^{\tau_L(u)-1}$, the objective is

$$\begin{aligned} J_{L,z}(\pi) &= \mathbb{E}_\pi \left[\sum_{h < \tau_L} \gamma^h r_z(s_h, a_h) \right] - \mathcal{C}_{L,z}(\pi), \\ \mathcal{C}_{L,z}(\pi) &:= \mathbb{E}_\pi \sum_{h < \tau_L} \gamma^h \alpha_z(s_h) \mathrm{KL}(\pi(\cdot \mid s_h) \| \pi_{\mathrm{ref}}(\cdot \mid s_h)). \end{aligned} \tag{2}$$

Bounded intermediate rewards are allowed in the Bellman results; any such non-IG returns must be included in $\mathcal{O}_z$ below using the same timing and discount. Horizons and costs count failures as well as accepted programs.

Assumption 3.6 (Temperature)**.** For all active states, $0 < \alpha_{\min} \leq \alpha_z(s) \leq \alpha_{\max} < \infty$. The bounds are uniform over time in recurrence statements. Lipschitz dependence on z is imposed separately for the ODE results, on a fixed finite tree.

4 CORE THEOREMS AND PROOF SKETCHES

The supporting results first identify the ideal policy and its information value. We then connect both to the implemented projected table in Theorem 4.12. All finite-round statements use the same reference, legal-action mask, temperature, and terminal-outcome space. Full proofs are in Section A.

4.1 REGULARIZED CONTROL AND REFERENCE-SUPPORTED PATHS

Proposition 4.1 (Soft Bellman fixed point and Gibbs policy)**.** *Fix L, z, assume Definition 3.1 and Assumption 3.6, and let $|r_z| \leq R_{\max}$. Set $V_{q,z}(s) = \alpha_z(s) \log \sum_a \pi_{\mathrm{ref}}(a \mid s) e^{q(s,a)/\alpha_z(s)}$, with $V = 0$ at terminal states, and $(\mathcal{T}_z q)(s,a) = r_z(s,a) + \gamma \mathbb{E}[V_{q,z}(s') \mid s, a]$. Then $\mathcal{T}_z$ is a γ-contraction and has the unique fixed point $q^*(z) = Q^*_{L,z}$. Its greedy policy is*

$$\pi^*_{L,z}(a \mid s) = \frac{\pi_{\mathrm{ref}}(a \mid s) e^{Q^*_{L,z}(s,a)/\alpha_z(s)}}{\sum_b \pi_{\mathrm{ref}}(b \mid s) e^{Q^*_{L,z}(s,b)/\alpha_z(s)}}. \tag{3}$$

Furthermore,

$$\|Q^*_{L,z}\|_\infty \leq \frac{R_{\max}}{1-\gamma}, \qquad \Delta_Q := \sup_{L,z,s,a,b} |Q^*_{L,z}(s,a) - Q^*_{L,z}(s,b)| \leq \frac{2R_{\max}}{1-\gamma}, \tag{4}$$

$$\pi^*_{L,z}(a \mid s) \geq \pi_{\mathrm{ref}}(a \mid s) e^{-\Delta_Q/\alpha_{\min}}. \tag{5}$$

Proof idea. Log-sum-exp is nonexpansive in the sup norm; the Gibbs variational identity supplies the greedy step. These are the regularized-control tools used throughout (Geist et al., 2019). Full proof: Section A.1.

For each fixed row, the KL penalty has the usual trust-region interpretation, with explicit degenerate and boundary cases in Proposition A.1.

Assumption 4.2 (Predictable support-preserving decoder)**.** Before round t, $L(t)$, G_t, and α_t are $\mathcal{F}_{t-1}$-measurable. Sampling uses all legal actions, without top-k or nucleus truncation, according to

$$\pi_t(a \mid s) = \frac{\pi_{\mathrm{ref}}(a \mid s) e^{G_t(s,a)/\alpha_t(s)}}{\sum_b \pi_{\mathrm{ref}}(b \mid s) e^{G_t(s,b)/\alpha_t(s)}}, \qquad |G_t(s,a)| \le B, \quad \alpha_t(s) \ge \alpha_{\min} > 0. \tag{6}$$

Here B and $\alpha_{\min}$ do not change with the horizon or round.

Lemma 4.3 (Completed-path floor)**.** *Under Definition 3.1 and Assumption 4.2, if $L(t) \ge \ell(u)$, the event $E_t(u)$ of emitting exactly the completed program u satisfies*

$$\mathbb{P}(E_t(u) \mid \mathcal{F}_{t-1}) \ge \kappa(u) := \pi_{\mathrm{ref}}(u) e^{-2B\ell(u)/\alpha_{\min}} > 0. \tag{7}$$

Multiplying the reference-relative token bounds, including `EOS`, proves the claim; see Section A.3. The implemented table will enforce this assumption directly, rather than through an assumed neural-policy tracking property.

A predictable conditional probability floor suffices for recurrence; independence is unnecessary (Lemma A.2).

Theorem 4.4 (Completed-program recurrence)**.** *Under Definition 3.1 and Assumption 4.2, let $L(t)$ be deterministic, nondecreasing, and divergent. Almost surely, every $u \in \mathcal{U}$ is emitted infinitely often, simultaneously over the countable language. If $t_u = \inf\{t : L(t) \ge \ell(u)\}$ and $T_u = \inf\{t \ge t_u : E_t(u)\}$, then*

$$\mathbb{E}[T_u] \le t_u + \kappa(u)^{-1}. \tag{8}$$

This follows from Lemmas A.2 and 4.3 and a countable intersection; see Section A.5. The potentially large $1/\kappa(u)$ is not a practical search-complexity claim.

Remark 4.5 (Bounded semantics). For a finite input set $\mathcal{I}$ and execution budget T, define $u \sim v$ when $E(u, i; T) = E(v, i; T)$ for all $i \in \mathcal{I}$. Each nonempty class inherits recurrence by choosing any of its valid members. This limited quotient consequence does not decide unrestricted program equivalence or remove representation multiplicity.

4.2 Information value and an IG-ablated comparator

Lemma 4.6 (Finite-ensemble entropy identity)**.** *For $w \in \Delta_K$, finite $\mathcal{Y}$, and component predictors p_m, let $\bar{p} = \sum_m w_m p_m$. Define*

$$w_m^+(y) = w_m p_m(y \mid x) / \bar{p}(y \mid x) \tag{9}$$

when $\bar{p}(y \mid x) > 0$, arbitrarily otherwise. Then

$$H(w) - \mathbb{E}_{Y \sim \bar{p}} H(w^+(Y)) = H(\bar{p}) - \sum_m w_m H(p_m) = \widehat{\mathrm{IG}}_{w,p}(x), \tag{10}$$

which lies in $[0, \min(\log K, \log |\mathcal{Y}|)]$. The conditional version requires the same mixture-predictive *observation law given the pre-observation state and query.*

This is $I(M; Y \mid x)$ computed from either marginal. It is an expectation identity, not monotone entropy decrease on every outcome. Full proof: Section A.6.

Assumption 4.7 (Surrogate error envelope)**.** For fixed L, z, $|\widetilde{\mathrm{IG}}_z(x) - I_z(x)| \le \delta_z(x)$ with finite nonnegative δ_z. Projection onto $[0, I_{\max}]$ also gives $|\widetilde{I}_z(x) - I_z(x)| \le \delta_z(x)$. When an estimated envelope is probabilistic, every conclusion using it is conditional on its simultaneous coverage event.

For the terminal convention of Definition 3.5, set

$$\begin{aligned} \mathcal{O}_z(\pi) &= \mathbb{E}_\pi[\Gamma_L(U) R_z^{\mathrm{other}}(U)], \quad \mathcal{D}_z(\pi) = \mathbb{E}_\pi[\Gamma_L(U) I_z(x(U))], \\ \widetilde{\mathcal{D}}_z(\pi) &= \mathbb{E}_\pi[\Gamma_L(U) \widetilde{I}_z(x(U))], \quad \mathcal{E}_z(\pi) = \mathbb{E}_\pi[\Gamma_L(U) \delta_z(x(U))]. \end{aligned} \tag{11}$$

Thus $J_{L,z} = \mathcal{O}_z + \lambda_I \widetilde{\mathcal{D}}_z - \mathcal{C}_{L,z}$ uses exactly the clipped reward. Write $J^{\mathrm{ex}}_{L,z} := \mathcal{O}_z + \lambda_I \mathcal{D}_z - \mathcal{C}_{L,z}$ for exact-information evaluation, not the objective optimized by a biased surrogate.

Proposition 4.8 (Information-directed objective)**.** *Assume Assumption 4.7 and $\lambda_I > 0$. In the pure-IG case, any maximizer $\pi_z^\dagger$ of $\lambda_I \widetilde{\mathcal{D}}_z - \mathcal{C}_{L,z}$ satisfies, for all π,*

$$\begin{aligned}\lambda_I \mathcal{D}_z(\pi_z^\dagger) - \mathcal{C}_{L,z}(\pi_z^\dagger) \geq \lambda_I \mathcal{D}_z(\pi) - \mathcal{C}_{L,z}(\pi) \\ - \lambda_I[\mathcal{E}_z(\pi) + \mathcal{E}_z(\pi_z^\dagger)].\end{aligned} \tag{12}$$

For composite rewards, add $\mathcal{O}_z$ to both sides and use the composite maximizer. In particular, choose any

$$\begin{aligned}&\pi_z^0 \in \arg\max_\pi \{\mathcal{O}_z(\pi) - \mathcal{C}_{L,z}(\pi)\}, \\ &\mathfrak{G}_z := (\mathcal{O}_z - \mathcal{C}_{L,z})(\pi_z^0) - (\mathcal{O}_z - \mathcal{C}_{L,z})(\pi_z^\star) \geq 0,\end{aligned} \tag{13}$$

where $\pi_z^\star = \pi^*_{L,z}$ optimizes the composite clipped objective. Then

$$\mathcal{D}_z(\pi_z^\star) \geq \mathcal{D}_z(\pi_z^0) + \mathfrak{G}_z/\lambda_I - \mathcal{E}_z(\pi_z^\star) - \mathcal{E}_z(\pi_z^0). \tag{14}$$

Proof idea. The finite policy space is compact and its objective continuous, so the comparator exists. Rearrange composite optimality against π_z^0, then bound the two surrogate errors. The comparator can depend on z and is not a new reference policy or a persistent endpoint. Full proof: Section A.7.

Corollary A.3 records the coarse-grained Gibbs formula when a separate terminal-law relaxation is justified. It is not used as a reduction of general discounted, state-dependent KL control.

4.3 One projected iterate, two certificates

Assumption 4.9 (Fixed-horizon regularity for dynamics)**.** Fix L. The fast iterate is projected to $\mathcal{Q}_L = [-C_Q, C_Q]^{\mathcal{X}_L}$ with $C_Q \geq R_{\max}/(1-\gamma)$. The slow state lies in a compact convex $\mathcal{Z}$. Rewards and temperatures are Lipschitz in z with constants L_r, L_α, and $h(q, z)$ is jointly Lipschitz. The reference and transition tree do not depend on z.

Assumption 4.10 (Two-time-scale recursion)**.** With Euclidean projections and the pre-generation convention of Definition 3.3,

$$q_{t+1} = \Pi_{\mathcal{Q}_L}\{q_t + a_t(\mathcal{T}_{z_t} q_t - q_t + M^q_{t+1})\}, \tag{15}$$

$$z_{t+1} = \Pi_{\mathcal{Z}}\{z_t + b_t(h(q_t, z_t) + M^z_{t+1} + \epsilon_t)\}. \tag{16}$$

The positive deterministic steps satisfy $\sum a_t = \sum b_t = \infty$, $\sum(a_t^2 + b_t^2) < \infty$, and $b_t/a_t \to 0$. Each M_{t+1} is $\mathcal{F}_t$-measurable, uniformly bounded, and conditionally centered given $\mathcal{F}_{t-1}$; $\epsilon_t \to 0$ a.s. Clipping a stochastic increment may introduce bias, which belongs in ϵ_t, not in the centered noise.

Definition 4.11 (Actual and ideal curricula)**.** For any $q \in \mathcal{Q}_L$ and z, let $\pi_{q,z}$ be (3) with $Q^*_{L,z}$ replaced by q. Define

$$\pi_t = \pi_{q_t, z_t}, \quad \pi_t^\star = \pi_{q^*(z_t), z_t}, \quad \mu_t = \mu_{\pi_t}, \quad \mu_t^\star = \mu_{\pi_t^\star}, \quad \varepsilon_t^{\mathrm{tr}} = \tfrac{1}{2}\|\mu_t - \mu_t^\star\|_1.$$

Thus $G_t = q_t$, $\alpha_t = \alpha_{z_t}$, and the decoder bound in Assumption 4.2 holds with $B = C_Q$.

Algorithm 1: Projected soft-Q endogenous curriculum. At round t: (1) freeze (q_t, z_t), the clipped score, and the objective; (2) form $\pi_t = \pi_{q_t, z_t}$; (3) sample a path and obtain execution/observation feedback; (4) form estimators of the Bellman and slow drifts; (5) install (15)–(16); (6) audit the pre-update residual against the frozen state. A finite model may compute the pending Bellman step before sampling ($M^q = 0$); it must still generate from q_t, not q_{t+1}. The finite Bayes solver in the experiment is a separate discrete specialization, not an application of the slow-SA theorem.

Lemmas A.4 and A.5 establish global exponential stability of the frozen flow and Lipschitz dependence of $q^*(z)$. Their proofs verify the regularity needed for the later two-time-scale limit.

Theorem 4.12 (Projected soft-Q coverage–information certificate)**.** *Fix L and a round-t objective satisfying Definitions 3.1 and 3.5 and Assumptions 3.6 and 4.7. Assume $\lambda_I > 0$ and $q_t \in [-C_Q, C_Q]^{\mathcal{X}_L}$*

with $C_Q > 0$. No convergence or time-scale assumption is needed for this finite-round statement. Use the actual and ideal curricula in Definition 4.11, and let $r_t = \|q_t - \mathcal{T}_{z_t} q_t\|_\infty$.

(i) Direct support. *For every completed u with $\ell(u) \le L$,*

$$\mu_t(u) \ge \kappa_t^{\mathrm{row}}(u) := \pi_{\mathrm{ref}}(u) \exp\left[- \sum_{k<\ell(u)} \frac{\max_b q_t(s_k, b) - q_t(s_k, a_k)}{\alpha_{z_t}(s_k)} \right] \tag{17}$$
$$\ge \pi_{\mathrm{ref}}(u) e^{-2C_Q \ell(u)/\alpha_{\min}}.$$

(ii) Residual-to-law tracking.

$$\varepsilon_t^{\mathrm{tr}} \le B_t^{\mathrm{TV}} := \tanh\left(\frac{L r_t}{\alpha_{\min}(1-\gamma)} \right). \tag{18}$$

(iii) Implemented information value.

$$\mathcal{D}_{z_t}(\pi_t) \ge \mathcal{D}_{z_t}(\pi_{z_t}^0) + \frac{\mathfrak{G}_{z_t}}{\lambda_I} - \mathcal{E}_{z_t}(\pi_t^\star) - \mathcal{E}_{z_t}(\pi_{z_t}^0) - I_{\max} \varepsilon_t^{\mathrm{tr}}. \tag{19}$$

Replacing $\varepsilon_t^{\mathrm{tr}}$ by B_t^{TV} gives the residual-composed lower bound.

Proof idea. Part (i) bounds each Gibbs normalizer by its largest exponent. For (ii), contraction gives $\|q_t - q^*\|_\infty \le r_t/(1-\gamma)$; token log-ratio errors accumulate over at most L actions. Part (iii) combines (14) with TV stability of the bounded nonnegative function $\Gamma_L I_{z_t}$. Full proofs: Sections A.11 and A.12. The exact residual is directly computable in the model-based case. With sampled targets one instead needs a confidence bound $r_t \le \widehat{r}_t + c_t$ before using (18).

Scope of the two certificates. Projection enforces support even with poor tracking. The row floor improves the *current* conditional probability bound; its observed inverse is not an adaptive future hitting-time guarantee. Information transfer can have a nonpositive RHS. The empirical section reports this rather than identifying positivity of the Gibbs density with efficient exploration.

Corollary 4.13 (Finite Bayesian entropy budget)**.** *Suppose a coherent finite Bayesian learner has belief w_t before round t, observation law $Y_t \mid \mathcal{F}_{t-1}, U_t \sim \sum_m w_t(m) p_m(\cdot \mid x(U_t))$, and updates w_{t+1} by* (9)*. Let*

$$A_t = \mathcal{D}_{z_t}(\pi_{z_t}^0) + \mathfrak{G}_{z_t}/\lambda_I, \qquad e_t = \mathcal{E}_{z_t}(\pi_t^\star) + \mathcal{E}_{z_t}(\pi_{z_t}^0) + I_{\max} \varepsilon_t^{\mathrm{tr}}.$$

With a fixed initial belief, integrable terms, and valid envelopes,

$$\mathbb{E}[H(w_t) - H(w_{t+1}) \mid \mathcal{F}_{t-1}] \ge A_t - e_t, \qquad \mathbb{E} H(w_{T+1}) \le H(w_1) - \sum_{t=1}^T \mathbb{E} A_t + \sum_{t=1}^T \mathbb{E} e_t. \tag{20}$$

Here $\mathbb{E}[H(w_t) - H(w_{t+1}) \mid \mathcal{F}_{t-1}] \ge \mathcal{D}_{z_t}(\pi_t)$ because $0 \le \Gamma_L \le 1$. This is not a pathwise entropy identity or a theorem for arbitrary SGD solvers. Full proof: Section A.13.

4.4 Direct residual certificates and identifiable learning

The path-law relaxation above is not needed to certify the information functional itself. Write $\mathcal{K}(\pi\|\sigma) = \mathbb{E}_\pi \sum_{h<\tau} \gamma^h \alpha(s_h) \mathrm{KL}(\pi_h \| \sigma_h)$ and $\Delta_t^{0,\mathrm{imp}} = J_{0,z_t}(\pi_{z_t}^0) - J_{0,z_t}(\pi_t)$, where $J_{0,z} = \mathcal{O}_z - \mathcal{C}_{L,z}$. The scalar $\Delta_t^{0,\mathrm{imp}}$ differs from the ideal-policy gap $\mathfrak{G}_{z_t}$ and the decoder table $G_t(s,a)$. Define $d_t = q_t - \mathcal{T}_{z_t} q_t$, its discounted path return $D_t^{\mathrm{res}}(u) = \sum_{h<\tau(u)} \gamma^h d_t(s_h, a_h)$, $\omega_t = \max_u D_t^{\mathrm{res}}(u) - \min_u D_t^{\mathrm{res}}(u)$, and $v_{t0} = \mathrm{TV}(\mu_t, \mu_{\pi_{z_t}^0})$. The contrast error is

$$E_{t0}^\Delta = \sum_u |\mu_t(u) - \mu_{\pi_{z_t}^0}(u)| \Gamma_L(u) \delta_{z_t}(x(u)).$$

Theorem 4.14 (Direct implemented information advantage)**.** *Under the finite-round hypotheses of Theorem 4.12, let $S_n = \sum_{j=0}^{n-1} \gamma^j$ and $H_L = \sum_{h=0}^{L-1} \gamma^h S_{L-h}^2$. A valid choice is*

$$\Xi_t = \min\left\{ 2 S_L r_t, \frac{H_L r_t^2}{2\alpha_{\min}} \right\}.$$

Then

$$\mathcal{D}_{z_t}(\pi_t) - \mathcal{D}_{z_t}(\pi^0_{z_t}) \geq \max\left\{ \frac{\Delta_t^{0,\mathrm{imp}} - \Xi_t}{\lambda_I}, \frac{\Delta_t^{0,\mathrm{imp}} + \mathcal{K}(\pi^0_{z_t}\|\pi_t) - \omega_t v_{t0}}{\lambda_I} \right\} - E^{\Delta}_{t0}. \tag{21}$$

Moreover $\Delta_t^{0,\mathrm{imp}} = \mathcal{K}(\pi_t\|\pi^0_{z_t}) \geq 0$. *The realized-row support floor remains unchanged for this same iterate.*

Proof idea and computation. The identity $J(\pi) = V_{q_t}(s_0) - \mathbb{E}_\pi D_t^{\mathrm{res}} - \mathcal{K}(\pi\|\pi_t)$ telescopes on the tree. At the optimal table, objective loss is a discounted policy KL; a log-moment bound makes it quadratic in residual error. The contrast charges the shared estimator error only through policy differences. Full proofs, sharper local bounds, and a positive construction are in Sections E to H. A positive RHS certifies comparator-relative advantage, not just positive information. The finite model and ablated comparator remain required, and these bounds have not been evaluated on v2/v3.

From support to identifiable learning. For a correctly specified finite Bayesian likelihood family, let $b_{ij}(u) = 1 - \sum_y \sqrt{p_i(y \mid u)p_j(y \mid u)}$ and let $k_t(u)$ be the current row floor. Define $A_{T,ij} = \sum_{t\leq T}\sum_u k_t(u)b_{ij}(u)$. Under a fixed true hypothesis i, with probability at least $1-\delta$, simultaneously over all T and $j \neq i$,

$$\frac{w_{T+1}(j)}{w_{T+1}(i)} \leq \frac{w_1(j)}{w_1(i)} \left(\frac{K-1}{\delta}\right)^2 e^{-2A_{T,ij}}.$$

The proof constructs a square-root likelihood-ratio supermartingale; Theorem J.1 and Corollary J.2 give all assumptions, the proof, finite-time bounds, and posterior consistency when every $A_{T,ij}$ diverges. This connects accumulated support to learning under identifiability, but neither asserts that IG is necessary nor compares whole-rollout learning rates against the raw reference.

4.5 Tracking limits and expanding horizons

Theorem 4.15 (Fixed-horizon two-time-scale limit)**.** *Under Assumptions 4.9 and 4.10, let* $\bar{h}(z) = h(q^*(z), z)$. *Then*

$$\|q_t - q^*(z_t)\|_\infty \longrightarrow 0 \quad \textit{a.s.} \tag{22}$$

The slow interpolation is an asymptotic pseudo-trajectory of

$$\dot{z} = \Gamma_{\mathcal{Z}}(z, \bar{h}(z)), \tag{23}$$

where $\Gamma_{\mathcal{Z}}$ *projects onto the tangent cone. Its limit set is compact, invariant, and internally chain transitive. If the semiflow admits a strict Lyapunov function* V *relative to its stationary set* $\mathcal{S}$ *and* $V(\mathcal{S})$ *has empty interior, then* $\mathrm{dist}(z_t, \mathcal{S}) \to 0$ *a.s.*

The projection, frozen-window tracking, and slow-limit arguments are proved in Section A.14; the Lyapunov qualification is the standard chain-transitivity condition (Benaïm, 1999). In particular,

$$r_t \leq (1+\gamma)\|q_t - q^*(z_t)\|_\infty \to 0,$$

so the tracking penalty in the certificate vanishes. Surrogate errors need their own control. A concrete projected-SGD drift satisfying the regularity assumptions is derived in Section B.

Theorem 4.16 (Epochwise consistency with expanding coverage)**.** *Let* $L_e \uparrow \infty$. *At each predictable epoch start* N_e, *suppose the chosen fixed-horizon recursion, if continued indefinitely at* L_e, *satisfies Theorem 4.15. End the epoch at the first available pre-generation state, after at least one round, with* $\|q - q^*_{L_e}(z)\|_\infty \leq \varepsilon_e$ *for a prescribed* $\varepsilon_e > 0$. *Every endpoint is a.s. finite. If the active decoders also satisfy Assumption 4.2 with common* $B, \alpha_{\min}$, *every* $u \in \mathcal{U}$ *recurs a.s. For the first eligible epoch start* τ_u *and* $T_u = \inf\{t \geq \tau_u : E_t(u)\}$,

$$\mathbb{E}[T_u - \tau_u \mid \mathcal{F}_{\tau_u - 1}] \leq \kappa(u)^{-1}. \tag{24}$$

Apply fixed-horizon tracking to each counterfactual continuation and then the conditional path-floor argument after the random start; see Section A.15. There is no common growing-dimensional ODE or uniform epoch runtime claim. Fixed-budget experimental epochs below are a diagnostic, not a verification of this stopping-time construction.

5 EXECUTABLE FINITE-MODEL EXPERIMENTS

Setup and comparisons. A bounded Boolean DSL supplies 192 programs spanning 16 two-input truth tables. An equality verifier compares each executed program with a hidden hypothesis; its observed bit is flipped independently with probability 0.02. The solver is an exact finite posterior. The online generator uses Monte Carlo IG, takes one projected model-based Bellman step per round, and samples from the pre-update table. Exact IG and backward solutions are frozen-state diagnostics, not inputs to online selection (Section C).

The v2 dataset has 17 configurations, 40 paired seeds per configuration, and 40 rounds. Targets use seed modulo 16; 95% intervals use 5,000 target-clustered bootstrap draws, without multiplicity adjustment. Full refresh takes 768,700 simulated outcomes per class per round for radius 0.02: an expensive model audit, not equal-compute evaluation. The later coverage-only dataset v3 is reported separately.

5.1 INFORMATION AND OBJECTIVE VALUE ON THE SAME BELIEF

For each sampled state, we evaluate π_t, $\pi_t^\star$, and $\pi_{z_t}^0$ under the same $\mathcal{O}_{z_t}$, information vector, and reference. In the baseline configuration their mean exact information values are respectively 0.03969, 0.04045, and 0.03040. The implemented policy therefore retains about 92% of the *mean* ideal information advantage; at $\alpha = 0.15$ the corresponding values are 0.04708, 0.05012, and 0.02637, retaining about 87%. These are ratios of aggregated functionals, not per-round lower bounds. The designated 9,600-round ordering audit reports no violation of $\widetilde{\mathcal{D}}(\pi^\star) \geq \widetilde{\mathcal{D}}(\pi^0)$. Exact-information ordering need not follow under bias: its reported 194 violations occur in the 1,600-round bias-0.25 arm.

To charge the cost of curriculum displacement, we also evaluate J^{ex}_{L,z_t}, not just information. Table 1 shows selected paired comparisons. Across the 17 configurations, ten intervals for $J^{\mathrm{ex}}(\pi_t) - J^{\mathrm{ex}}(\pi^0)$ are positive, five include zero, and two are negative. The ideal surrogate-optimal policy has the largest reported exact-evaluation mean among the audited policies in all 17 configurations; this is an observation, not a theorem about biased optimization.

Table 1: **Same-snapshot exact-objective gain** $J^{\mathrm{ex}}_{L,z_t}(\pi_t) - J^{\mathrm{ex}}_{L,z_t}(\pi^0_{z_t})$ (v2). Positive values favor the implemented generator. Stress magnitudes are configuration parameters, not percentage errors.

Configuration	Mean gain	95% target-clustered interval
Pure information	$+0.00257$	$[+0.00136, +0.00393]$
Pure information, $\alpha = 0.15$	$+0.00356$	$[+0.00169, +0.00564]$
Composite, 20 Q-steps/round	$+0.00217$	$[+0.00127, +0.00303]$
Reference misalignment 0.6	-0.00099	$[-0.00125, -0.00073]$
Injected bias 0.25	-0.00316	$[-0.00354, -0.00273]$

5.2 LEARNING CURVES AND THE OBJECTIVE TRADE-OFF

Under the matched pure-information objective, both the varying program-cost term and transient shaping are disabled. The IG-ablated optimum then coincides with the original reference. Separate rollouts favor the online projected curriculum on target success, capped time to target, semantic coverage, and concentration (Table 2). The entropy-AUC difference is -4.29, but its target-clustered interval is borderline: $[-8.52, -0.03]$ against the reference and $[-8.36, +0.11]$ against the numerically equivalent ablated comparator under different bootstrap resamples. We therefore do not rely on its significance label.

Under the composite objective, the raw reference has better learning-only point estimates in the baseline setting. The factorial cost/shaping controls associate most of this difference with the preference for cheaper programs. This is a trade-off to report, not a contradiction of composite optimality. Snapshot optimality and rollout learning are different comparisons.

Table 2: **Pure-information rollout effects** (online projected minus raw reference, v2). Semantic coverage is the fraction of 16 truth tables visited. Time is capped at the common observation budget plus one, not an uncensored mean hitting time.

Metric	Mean difference	95% target-clustered interval
Target success (higher is better)	$+0.125$	$[+0.025, +0.250]$
Capped time to target (lower)	-4.20	$[-7.00, -1.54]$
Semantic coverage (higher)	$+0.052$	$[+0.029, +0.074]$
Mean maximum class share (lower)	-0.020	$[-0.027, -0.013]$

5.3 What the certificates do and do not certify

Define R_t^{exact} as the RHS of (19) using audited exact TV, and R_t^{comp} by substituting B_t^{TV}. Their non-vacuity criterion is strictly positive RHS, since $\mathcal{D}_z \geq 0$ already holds universally. In v2 the baseline has $R_t^{\mathrm{exact}} > 0$ in 288/1600 rounds, increasing to 658/1600 at 20 Q-steps per round. The residual-composed RHS remains nonpositive in all 1600 rounds of every tested configuration. Thus a tighter tracker improves exact-information audits without making the tested residual-derived lower bound useful. This identifies a limiting relaxation, not a violation of the theorem or a proof that no better tracking budget could help.

Bias exposes a second distinction. With configured perturbation magnitudes $0, 0.05, 0.10, 0.25$, exact-TV positive counts are respectively $288, 12, 1, 0$ per 1600 rounds. Only the largest perturbation yields a clearly negative objective gain among these arms. The certificate loses positive content before the measured objective effect is conclusively negative; absence of statistical significance at smaller perturbations is not evidence of equality.

The original box floor is likewise very small. In the separate v3 coverage audit, the baseline's reported row-floor median is 5.47×10^{-4} and the pure-IG median is 2.21×10^{-3}, compared with a box-floor median of order 10^{-55} at $C_Q = 4$, $\alpha = 0.4$. The seven explicitly enumerated v3 configurations have reported row floors above the numerical diagnostic threshold 10^{-12} for every program in every audited round; none violates (17). Full per-configuration values are in Table 5. These are stronger *current* probability bounds, not certified adaptive future hitting times or a general temperature-invariance theorem.

6 Limitations

The guarantees concern the specified decoder and finite likelihood model, not arbitrary neural optimization or calibration. Support alone does not imply rapid discovery. Model-based audits are expensive, and fixed-budget epochs do not verify the SA or stopping-time hypotheses. Exploratory intervals, unbalanced target repetitions, and replication limits are detailed in Section C.7.

7 Conclusion

One projected soft-Q state supports both program non-extinction and information certificates. Direct residual identities sharpen the comparison, and identifiable likelihoods connect accumulated support to posterior learning. The finite experiments show conditional gains, not a useful value of the original residual-composed bound. Evaluating the direct refinements on those records and comparing complete adaptive learning processes remain separate tasks.

REFERENCES

Matej Balog, Alexander L. Gaunt, Marc Brockschmidt, Sebastian Nowozin, and Daniel Tarlow. Deepcoder: Learning to write programs. In *International Conference on Learning Representations*, 2017.

Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Séminaire de Probabilités XXXIII*, volume 1709 of *Lecture Notes in Mathematics*, pp. 1–68. Springer, 1999.

Justin Yang Chae, Md Tanvirul Alam, and Nidhi Rastogi. Towards understanding self-play for LLM reasoning. *arXiv preprint arXiv:2510.27072*, 2025.

Bei Chen, Fengji Zhang, Anh Nguyen, Daoguang Zan, Zeqi Lin, Jian-Guang Lou, and Weizhu Chen. Codet: Code generation with generated tests. In *International Conference on Learning Representations*, 2023.

Xiaoyin Chen, Jiarui Lu, Minsu Kim, Dinghuai Zhang, Jian Tang, Alexandre Piché, Nicolas Gontier, Yoshua Bengio, and Ehsan Kamalloo. Self-evolving curriculum for LLM reasoning. *arXiv preprint arXiv:2505.14970*, 2025.

Xinyun Chen, Chang Liu, and Dawn Song. Execution-guided neural program synthesis. In *International Conference on Learning Representations*, 2019.

Jacob Devlin, Jonathan Uesato, Surya Bhupatiraju, Rishabh Singh, Abdel-rahman Mohamed, and Pushmeet Kohli. Robustfill: Neural program learning under noisy i/o. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 990–998, 2017.

Kevin Ellis, Catherine Wong, Maxwell Nye, Mathias Sablé-Meyer, Luc Cary, Lucas Morales, Luke Hewitt, Armando Solar-Lezama, and Joshua B. Tenenbaum. Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep bayesian program learning. *Philosophical Transactions of the Royal Society A*, 381(2251):20220050, 2023. doi: 10.1098/rsta.2022.0050.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 1050–1059, 2016.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2160–2169, 2019.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1861–1870, 2018.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven C. H. Hoi. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

Sirin Nitinawarat, George K. Atia, and Venugopal V. Veeravalli. Controlled sensing for multi-hypothesis testing. *IEEE Transactions on Automatic Control*, 58(10):2451–2464, 2013. doi: 10.1109/TAC.2013.2261188.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, 2018.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Epistemic neural networks. *arXiv preprint arXiv:2107.08924*, 2021.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

Rémy Portelas, Cédric Colas, Lilian Weng, Katja Hofmann, and Pierre-Yves Oudeyer. Automatic curriculum learning for deep reinforcement learning: A short survey. *arXiv preprint arXiv:2003.04664*, 2020.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1889–1897, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Sainbayar Sukhbaatar, Zeming Lin, Ilya Kostrikov, Gabriel Synnaeve, Arthur Szlam, and Rob Fergus. Intrinsic motivation and automatic curricula via asymmetric self-play. In *International Conference on Learning Representations*, 2018.

Vatsal Venkatkrishna, Indraneil Paul, and Iryna Gurevych. Aletheia: What makes RLVR for code verifiers tick? *arXiv preprint arXiv:2601.12186*, 2026.

Xumeng Wen, Zihan Liu, Shun Zheng, Zhijian Xu, Shengyu Ye, Zhirong Wu, Xiao Liang, Yang Wang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*, 2025.

A Complete Proofs

This appendix gives the complete supporting proofs. The new coupled certificate is proved after the fixed-horizon stability lemmas and before the tracking limit, preserving the dependency order in the main text.

A.1 Proof of Proposition 4.1

Proof. We prove the result in five steps. Throughout the proof, the horizon L and slow state z are fixed. This is important: no statement below is taken over an expanding state space.

Step 1: terminal convention and Bellman operator. Let ∂ denote an absorbing terminal state. If action a completes the program, e.g. $a = \texttt{EOS}$, or if appending a reaches the horizon, we write $s \oplus a = \partial$ and define the continuation value at ∂ to be zero. For Bellman continuation only, terminal endpoints share ∂; the outcome law on Ω_L still records each distinct terminal path. Otherwise $s \oplus a$ is the next valid prefix. For any bounded function $Q : \mathcal{X}_L \to \mathbb{R}$, define its soft value at a nonterminal prefix s by

$$V_Q(s) := \alpha_z(s) \log \sum_{a \in \mathcal{A}(s)} \pi_{\mathrm{ref}}(a \mid s) \exp(Q(s,a)/\alpha_z(s)), \tag{25}$$

and set $V_Q(\partial) = 0$. The soft Bellman operator is

$$(\mathcal{T}_{L,z} Q)(s,a) := r_z(s,a) + \gamma V_Q(s \oplus a), \qquad (s,a) \in \mathcal{X}_L. \tag{26}$$

The space $(\mathbb{R}^{\mathcal{X}_L}, \|\cdot\|_\infty)$ is complete because $\mathcal{X}_L$ is finite: Σ is finite, L is fixed, and every $\mathcal{A}(s) \subseteq \Sigma \cup \{\texttt{EOS}\}$ is finite.

Step 2: the soft value is a variational maximum. For a fixed prefix s and vector $q \in \mathbb{R}^{\mathcal{A}(s)}$, define

$$\Phi_s(p; q) := \sum_{a \in \mathcal{A}(s)} p(a) q(a) - \alpha_z(s) \mathrm{KL}\big(p \,\|\, \pi_{\mathrm{ref}}(\cdot \mid s)\big), \qquad p \in \Delta(\mathcal{A}(s)).$$

Because Definition 3.1 gives $\pi_{\mathrm{ref}}(a \mid s) > 0$ for every valid action, the KL term is finite for all $p \in \Delta(\mathcal{A}(s))$. Let

$$g_q(a) := \frac{\pi_{\mathrm{ref}}(a \mid s) \exp(q(a)/\alpha_z(s))}{\sum_{b \in \mathcal{A}(s)} \pi_{\mathrm{ref}}(b \mid s) \exp(q(b)/\alpha_z(s))}. \tag{27}$$

A direct rearrangement gives

$$\Phi_s(p; q) = \alpha_z(s) \log \sum_{b \in \mathcal{A}(s)} \pi_{\mathrm{ref}}(b \mid s) e^{q(b)/\alpha_z(s)} - \alpha_z(s) \mathrm{KL}(p \| g_q).$$

Since $\mathrm{KL}(p \| g_q) \geq 0$, equality is attained uniquely at $p = g_q$. Therefore

$$V_Q(s) = \max_{p \in \Delta(\mathcal{A}(s))} \left\{ \sum_a p(a) Q(s,a) - \alpha_z(s) \mathrm{KL}(p \| \pi_{\mathrm{ref}}(\cdot \mid s)) \right\}, \tag{28}$$

and the unique maximizer is the reference-weighted Gibbs distribution in (27). The uniqueness follows from strict convexity of $p \mapsto \mathrm{KL}(p \| \pi_{\mathrm{ref}})$ on the finite simplex relative to a full-support reference.

Step 3: contraction and fixed point. We first show that $Q \mapsto V_Q$ is 1-Lipschitz in sup norm. Let $\varepsilon := \|Q - Q'\|_\infty$. Then for every valid prefix s and every $a \in \mathcal{A}(s)$, $Q(s,a) \leq Q'(s,a) + \varepsilon$. Substituting this into (25) gives $V_Q(s) \leq V_{Q'}(s) + \varepsilon$. Reversing Q and Q' yields

$$|V_Q(s) - V_{Q'}(s)| \leq \|Q - Q'\|_\infty. \tag{29}$$

Using (26), for any $(s,a) \in \mathcal{X}_L$,

$$|(\mathcal{T}_{L,z} Q)(s,a) - (\mathcal{T}_{L,z} Q')(s,a)| \leq \gamma \|Q - Q'\|_\infty.$$

Thus $\mathcal{T}_{L,z}$ is a γ-contraction on the complete metric space $(\mathbb{R}^{\mathcal{X}_L}, \|\cdot\|_\infty)$. By Banach's fixed-point theorem, there exists a unique bounded fixed point $Q^*_{L,z} = \mathcal{T}_{L,z} Q^*_{L,z}$. The regularized dynamic-programming optimality equation is exactly this fixed-point equation, and the maximizing policy at each prefix is the Gibbs policy obtained in Step 2. Substituting $Q = Q^*_{L,z}$ into (27) proves (3).

Step 4: uniform value and span bounds. Let $M := \|Q^*_{L,z}\|_\infty$. From (25), $V_{Q^*_{L,z}}(s)$ lies between $\min_a Q^*_{L,z}(s,a)$ and $\max_a Q^*_{L,z}(s,a)$, because it is a log-sum-exp with normalized positive weights. Hence $|V_{Q^*_{L,z}}(s)| \le M$ for every s. The fixed-point equation gives

$$|Q^*_{L,z}(s,a)| \le R_{\max} + \gamma M.$$

Taking the supremum over (s,a) yields $M \le R_{\max} + \gamma M$, hence

$$\|Q^*_{L,z}\|_\infty \le \frac{R_{\max}}{1-\gamma}.$$

Consequently, for any s, a, b,

$$|Q^*_{L,z}(s,a) - Q^*_{L,z}(s,b)| \le 2\|Q^*_{L,z}\|_\infty \le \frac{2R_{\max}}{1-\gamma}.$$

Taking the supremum over L, z, s, a, b proves (4). The bound is uniform because $R_{\max}$ and γ are uniform across horizons and slow states.

Step 5: lower bound relative to the reference. Fix s and $a \in \mathcal{A}(s)$, and write $\alpha = \alpha_z(s)$. From (3),

$$\begin{aligned}
\pi^*_{L,z}(a \mid s) &= \frac{\pi_{\mathrm{ref}}(a \mid s) e^{Q^*(s,a)/\alpha}}{\sum_b \pi_{\mathrm{ref}}(b \mid s) e^{Q^*(s,b)/\alpha}} \\
&\ge \frac{\pi_{\mathrm{ref}}(a \mid s) e^{Q^*(s,a)/\alpha}}{e^{\max_b Q^*(s,b)/\alpha}} \\
&= \pi_{\mathrm{ref}}(a \mid s) \exp\left(-\frac{\max_b Q^*(s,b) - Q^*(s,a)}{\alpha}\right) \\
&\ge \pi_{\mathrm{ref}}(a \mid s) \exp\left(-\frac{\Delta_Q}{\alpha_z(s)}\right) \ge \pi_{\mathrm{ref}}(a \mid s) \exp\left(-\frac{\Delta_Q}{\alpha_{\min}}\right),
\end{aligned}$$

where the denominator bound used $\sum_b \pi_{\mathrm{ref}}(b \mid s) = 1$, and the final step used $\alpha_z(s) \ge \alpha_{\min}$. This proves (5). □

A.2 Proof of Proposition A.1

Proposition A.1 (Local KL penalty/trust-region duality). *Fix $A = \mathcal{A}(s)$, $\rho = \pi_{\mathrm{ref}}(\cdot \mid s) > 0$, and $Q \in \mathbb{R}^A$. For $\alpha > 0$, the unique maximizer of $\langle p, Q\rangle - \alpha \mathrm{KL}(p\|\rho)$ is*

$$p_\alpha(a) = \frac{\rho(a) e^{Q(a)/\alpha}}{\sum_b \rho(b) e^{Q(b)/\alpha}}. \tag{30}$$

If Q is nonconstant and $G = \arg\max Q$, $\mathrm{KL}(p_\alpha\|\rho)$ decreases continuously from $\beta_\star = -\log \rho(G)$ to zero as α increases. Each $\beta \in (0, \beta_\star)$ therefore selects a unique penalized solution that also uniquely solves $\max\{\langle p, Q\rangle : \mathrm{KL}(p\|\rho) \le \beta\}$. At $\beta = 0$ the solution is ρ; at $\beta \ge \beta_\star$ the optimum is attained on the feasible part of the greedy face. Constant Q is a degenerate case, not a one-to-one dual path.

Proof. We prove each part explicitly because the boundary cases are important for the interpretation of adaptive KL schedules.

Part (a): penalized optimum. For $p \in \Delta(A)$ define

$$F_\alpha(p) := \langle p, Q\rangle - \alpha \mathrm{KL}(p\|\rho).$$

Since $\rho(a) > 0$ for every $a \in A$, the KL term is finite on the whole simplex. Moreover, $p \mapsto \mathrm{KL}(p\|\rho)$ is strictly convex on the finite simplex, so F_α is strictly concave for every $\alpha > 0$. Hence the maximizer is unique.

Let

$$Z_\alpha := \sum_{b \in A} \rho(b) e^{Q(b)/\alpha}, \qquad g_\alpha(a) := \rho(a) e^{Q(a)/\alpha} / Z_\alpha.$$

A direct rearrangement gives, for every $p \in \Delta(A)$,

$$F_\alpha(p) = \alpha \log Z_\alpha - \alpha \mathrm{KL}(p\|g_\alpha).$$

Thus $F_\alpha(p)$ is maximized uniquely at $p = g_\alpha$, proving (30).

Part (b): monotone budget curve. It is cleaner to use the inverse temperature $\eta := 1/\alpha$. Write

$$p_\eta(a) = \rho(a)e^{\eta Q(a)}/Z(\eta), \qquad Z(\eta) = \sum_{a\in A} \rho(a)e^{\eta Q(a)}.$$

The induced KL budget is

$$\beta(\eta) := \mathrm{KL}(p_\eta\|\rho) = \eta\,\mathbb{E}_{p_\eta}[Q] - \log Z(\eta). \tag{31}$$

Since A is finite, Z is smooth. Differentiating (31) and using $\frac{d}{d\eta}\log Z(\eta) = \mathbb{E}_{p_\eta}[Q]$ gives

$$\beta'(\eta) = \eta\,\frac{d}{d\eta}\mathbb{E}_{p_\eta}[Q] = \eta\,\mathrm{Var}_{p_\eta}(Q). \tag{32}$$

If Q is nonconstant, then p_η has full support for every finite η, so $\mathrm{Var}_{p_\eta}(Q) > 0$ for every $\eta > 0$. Hence $\beta(\eta)$ is strictly increasing in η, and therefore $\beta(\alpha)$ is strictly decreasing in α.

The limit $\eta \downarrow 0$ gives $p_\eta \to \rho$ and hence $\beta \to 0$. This is the same as $\alpha \to \infty$. For $\eta \to \infty$, the Gibbs distribution concentrates on the maximizing set G and, inside G, remains proportional to the reference:

$$p_\eta(a) \to \rho_G(a) := \frac{\rho(a)\mathbf{1}\{a \in G\}}{\rho(G)}.$$

Therefore

$$\lim_{\eta\to\infty} \mathrm{KL}(p_\eta\|\rho) = \mathrm{KL}(\rho_G\|\rho) = \sum_{a\in G} \frac{\rho(a)}{\rho(G)} \log \frac{1}{\rho(G)} = -\log \rho(G) = \beta_\star(Q,\rho),$$

which proves the stated limits.

Part (c): active trust-region duality. Fix $\beta \in (0, \beta_\star)$ and let $\alpha = \alpha(\beta)$ be the unique value from Part (b), so that $\mathrm{KL}(p_\alpha\|\rho) = \beta$. Since p_α is feasible for the trust-region problem, it remains only to show that no other feasible distribution has larger linear value.

By Part (a), p_α maximizes the penalized objective. Hence, for every $p \in \Delta(A)$,

$$\langle p_\alpha, Q\rangle - \alpha\mathrm{KL}(p_\alpha\|\rho) \geq \langle p, Q\rangle - \alpha\mathrm{KL}(p\|\rho). \tag{33}$$

If p is feasible for the trust-region problem, then $\mathrm{KL}(p\|\rho) \leq \beta = \mathrm{KL}(p_\alpha\|\rho)$. Combining this with (33) gives

$$\langle p_\alpha, Q\rangle - \alpha\beta \geq \langle p, Q\rangle - \alpha\mathrm{KL}(p\|\rho) \geq \langle p, Q\rangle - \alpha\beta.$$

Therefore $\langle p_\alpha, Q\rangle \geq \langle p, Q\rangle$ for every feasible p, so p_α solves the constrained problem.

The optimizer is unique. If a feasible p also satisfies $\langle p, Q\rangle = \langle p_\alpha, Q\rangle$, then both inequalities above must be equalities. In particular, p also maximizes the strictly concave penalized objective F_α, whose unique maximizer is p_α by Part (a). Hence $p = p_\alpha$. This proves the active-budget duality without needing to assume a priori that the constrained optimizer lies in the relative interior of the simplex.

Part (d): boundary cases. For $\beta = 0$, the constraint $\mathrm{KL}(p\|\rho) \leq 0$ forces $p = \rho$, since KL divergence is nonnegative and vanishes only at equality. This agrees with the limiting penalized optimizer $p_\alpha \to \rho$ as $\alpha \to \infty$.

For $\beta \geq \beta_\star$, the distribution ρ_G is feasible and satisfies $\langle \rho_G, Q\rangle = \max_a Q(a)$. Therefore the constrained problem attains the unconstrained linear optimum. Any optimizer must be supported on G; otherwise it assigns positive mass to an action with strictly smaller value and obtains expected score strictly below $\max_a Q(a)$. Conversely, any distribution supported on G and satisfying $\mathrm{KL}(p\|\rho) \leq \beta$ attains the same optimal value. These greedy-face optimizers are obtained by the penalized family only in the zero-temperature limit $\alpha \downarrow 0$ in general, not at finite positive α.

Finally, if Q is constant then $G = A$ and $\beta_\star = 0$. The penalized objective reduces to a constant minus $\alpha\mathrm{KL}(p\|\rho)$, so its unique maximizer is ρ for every $\alpha > 0$. In contrast, the constrained linear objective is constant on the feasible set, so every feasible p is optimal. This is the degenerate case in which there is no nontrivial active trust-region path. □

A.3 Proof of Lemma 4.3

Proof. Fix a completed program u, write $\ell = \ell(u)$, and fix t with $L(t) \geq \ell$. Since π_t is $\mathcal{F}_{t-1}$-measurable, we may condition on an arbitrary realization of $\mathcal{F}_{t-1}$ and treat π_t as deterministic during the time-t episode. We first prove a one-step lower bound. For any valid prefix s and action $a \in \mathcal{A}(s)$, set $\alpha = \alpha_t(s)$ and $G_b = G_t(s, b)$. From (6),

$$\begin{aligned}\sum_{b\in\mathcal{A}(s)} \pi_{\mathrm{ref}}(b \mid s)\exp(G_b/\alpha) &\leq \exp(B/\alpha) \sum_{b\in\mathcal{A}(s)} \pi_{\mathrm{ref}}(b \mid s)\\ &= \exp(B/\alpha),\end{aligned}$$

and the numerator is at least $\pi_{\mathrm{ref}}(a \mid s)\exp(-B/\alpha)$. Hence

$$\pi_t(a \mid s) \geq \pi_{\mathrm{ref}}(a \mid s)\exp(-2B/\alpha) \geq \pi_{\mathrm{ref}}(a \mid s)\exp(-2B/\alpha_{\min}), \tag{34}$$

where the last inequality uses $\alpha \geq \alpha_{\min}$.

Now apply the chain rule for the autoregressive sampling process. Conditional on $\mathcal{F}_{t-1}$, the probability of the exact path u is

$$\begin{aligned}\mathbb{P}(E_t(u) \mid \mathcal{F}_{t-1}) &= \prod_{k=1}^{\ell} \pi_t(a_k \mid s_{k-1})\\ &\geq \prod_{k=1}^{\ell} \pi_{\mathrm{ref}}(a_k \mid s_{k-1})\exp(-2B\ell/\alpha_{\min})\\ &= \pi_{\mathrm{ref}}(u)\exp(-2B\ell(u)/\alpha_{\min}).\end{aligned}$$

This proves (7). Positivity follows from Definition 3.1 and finite $\ell(u)$. □

A.4 Proof of Lemma A.2

Lemma A.2 (Conditional recurrence criterion)**.** *If $A_t \in \mathcal{F}_t$ and, for deterministic $t_0, \kappa > 0$,*

$$\mathbb{P}(A_t \mid \mathcal{F}_{t-1}) \geq \kappa \quad (t \geq t_0) \quad \textit{a.s.}, \tag{35}$$

then A_t occurs infinitely often a.s. For $T = \inf\{t \geq t_0 : A_t\}$,

$$\mathbb{E}[T] \leq t_0 + \kappa^{-1}. \tag{36}$$

Proof. The proof uses only conditional probabilities and does not assume independence. For $n \geq 0$, let

$$F_n := \bigcap_{j=0}^{n} A_{t_0+j}^c$$

be the event of no success in the first $n+1$ trials after t_0. We prove by induction that

$$\mathbb{P}(F_n \mid \mathcal{F}_{t_0-1}) \leq (1-\kappa)^{n+1} \qquad \text{a.s.} \tag{37}$$

For $n = 0$ this is immediate from (35). If the claim holds for $n-1$, then using $F_{n-1} \in \mathcal{F}_{t_0+n-1}$ and the tower property,

$$\begin{aligned}\mathbb{P}(F_n \mid \mathcal{F}_{t_0-1}) &= \mathbb{E}\big[\mathbf{1}_{F_{n-1}}\mathbb{P}(A_{t_0+n}^c \mid \mathcal{F}_{t_0+n-1}) \mid \mathcal{F}_{t_0-1}\big]\\ &\leq (1-\kappa)\mathbb{P}(F_{n-1} \mid \mathcal{F}_{t_0-1}) \leq (1-\kappa)^{n+1}.\end{aligned}$$

This proves (37). Letting $n \to \infty$ gives

$$\mathbb{P}(T = \infty \mid \mathcal{F}_{t_0-1}) = 0 \quad \text{a.s.}$$

so $T < \infty$ a.s.

To show infinitely many occurrences, fix any deterministic $N \geq t_0$ and apply the same argument with N in place of t_0; with probability one there is at least one occurrence after N. The event that

only finitely many A_t occur is contained in the countable union, over $N \in \mathbb{N}$, of the event that no A_t occurs after N. Each such event has probability zero, hence A_t occurs infinitely often almost surely.

For the expectation, use the tail-sum formula for the nonnegative integer-valued random variable $T - t_0$:

$$\begin{aligned}\mathbb{E}[T - t_0] &= \sum_{n=0}^{\infty} \mathbb{P}(T - t_0 > n) \\ &\leq \sum_{n=0}^{\infty} (1-\kappa)^{n+1} = \frac{1-\kappa}{\kappa} \leq \kappa^{-1}.\end{aligned}$$

Adding t_0 yields (36). □

A.5 Proof of Theorem 4.4

Proof. Fix a completed valid program u. Since $L(t)$ is nondecreasing and tends to infinity, $t_u < \infty$. For every $t \geq t_u$, Lemma 4.3 gives

$$\mathbb{P}(E_t(u) \mid \mathcal{F}_{t-1}) \geq \kappa(u) \qquad \text{a.s.}$$

Apply Lemma A.2 with $A_t = E_t(u)$, $t_0 = t_u$, and $\kappa = \kappa(u)$. This proves the single-program recurrence claim and (8).

Finally, $\mathcal{U}$ is countable because it is a subset of $(\Sigma \cup \{\texttt{EOS}\})^*$ over a finite alphabet. For each fixed $u \in \mathcal{U}$, the event $C_u := \{E_t(u) \text{ occurs infinitely often}\}$ has probability one. Therefore

$$\mathbb{P}\left(\bigcap_{u \in \mathcal{U}} C_u\right) = 1 - \mathbb{P}\left(\bigcup_{u \in \mathcal{U}} C_u^c\right) \geq 1 - \sum_{u \in \mathcal{U}} \mathbb{P}(C_u^c) = 1,$$

which proves the simultaneous recurrence claim. □

A.6 Proof of Lemma 4.6

Proof. We prove the identity at a fixed query x; the conditional statement then follows by applying the same deterministic identity after conditioning on $\mathcal{F}_{t-1}$.

Step 1: construct the joint epistemic-predictive distribution. Define a joint distribution on $(M, Y) \in [K] \times \mathcal{Y}$ by

$$\mathbb{P}(M = m, Y = y) := w_m p_m(y \mid x). \tag{38}$$

Its marginal on Y is exactly the mixture predictive:

$$\mathbb{P}(Y = y) = \sum_{m=1}^{K} w_m p_m(y \mid x) = \bar{p}(y \mid x).$$

For every y such that $\bar{p}(y \mid x) > 0$, Bayes' rule gives

$$\mathbb{P}(M = m \mid Y = y) = \frac{w_m p_m(y \mid x)}{\bar{p}(y \mid x)} = w_m^+(y),$$

which is precisely (9). Values of $w^+(y)$ on outcomes with $\bar{p}(y \mid x) = 0$ do not affect any expectation under $Y \sim \bar{p}(\cdot \mid x)$.

Step 2: write mutual information through the predictive entropy. The mutual information between the epistemic index M and the answer Y under (38) is

$$\begin{aligned}I(M; Y \mid x) &= H(Y \mid x) - H(Y \mid M, x) \\ &= H(\bar{p}(\cdot \mid x)) - \sum_{m=1}^{K} w_m H(p_m(\cdot \mid x)) = \widehat{\mathrm{IG}}_{w,p}(x).\end{aligned} \tag{39}$$

The first equality is the definition of mutual information, and the second uses that $Y \mid (M = m, x)$ has distribution $p_m(\cdot \mid x)$.

Step 3: write the same mutual information through posterior entropy reduction. The same mutual information also satisfies

$$\begin{aligned} I(M;Y \mid x) &= H(M) - H(M \mid Y, x) \\ &= H(w) - \sum_{y \in \mathcal{Y}} \bar{p}(y \mid x) H(w^+(y)) = H(w) - \mathbb{E}_Y[H(w^+(Y))]. \end{aligned} \tag{40}$$

Here $H(M) = H(w)$, and the conditional distribution of M given $Y = y$ is $w^+(y)$ for all y with positive mixture probability. Combining (39) and (40) proves (10).

Step 4: nonnegativity, boundedness, and conditional form. Because mutual information is nonnegative and bounded by each marginal entropy,

$$\begin{aligned} 0 \le I(M;Y \mid x) \le H(M) = H(w) \le \log K, \\ I(M;Y \mid x) \le H(Y \mid x) = H(\bar{p}(\cdot \mid x)) \le \log |\mathcal{Y}|. \end{aligned}$$

Together with (10), this proves the stated bounds.

For the conditional statement, condition on $\mathcal{F}_{t-1}$. The objects w_t, $p_{t,m}$, and x_t become fixed quantities, and Y_t has conditional law $\bar{p}_t(\cdot \mid x_t)$. Applying the already-proved deterministic identity to these fixed quantities yields the conditional identity stated above. This proves the lemma. □

A.7 Proof of Proposition 4.8

Proof. We prove the four parts separately. Throughout the proof, z and L are fixed, so all expectations are taken over only the randomness of the generator trajectory and, where explicitly indicated, the answer sampled from the finite predictive mixture.

Clipping and common objective. Write $I_z = \widehat{\mathrm{IG}}_z$ and let $\widetilde{I}_z$ be the once-clipped score in Definition 3.4. Since $I_z(x) \in [0, I_{\max}]$, projection onto that interval gives

$$|\widetilde{I}_z(x) - I_z(x)| \le |\widetilde{\mathrm{IG}}_z(x) - I_z(x)| \le \delta_z(x).$$

Every functional below uses the same clipped score and terminal law from (11). In particular, clipping a different copy only during the audit would not justify this proof.

Part (a): entropy reduction averaged over the generated curriculum. For a fixed terminal outcome $u \in \Omega_L$, Lemma 4.6 applied to the task input $x(u)$ gives

$$\widehat{\mathrm{IG}}_z(x(u)) = \mathbb{E}_{Y \sim \bar{p}_z(\cdot \mid x(u))}\big[H(w_z) - H(w_z^+(Y))\big]. \tag{41}$$

Multiplying both sides by the nonnegative scalar $\Gamma_L(u)$ and averaging over U_π yields

$$\begin{aligned} \mathcal{D}_z(\pi) &= \mathbb{E}_\pi\big[\Gamma_L(U_\pi)\widehat{\mathrm{IG}}_z(x(U_\pi))\big] \\ &= \mathbb{E}_{U_\pi, Y}\big[\Gamma_L(U_\pi)\big(H(w_z) - H(w_z^+(Y))\big)\big], \end{aligned}$$

where the second expectation samples U_π from the generator and then samples $Y \sim \bar{p}_z(\cdot \mid x(U_\pi))$. This proves the entropy-reduction identity claimed in the proposition.

Part (b): pure information-design objective. If $R_z(u) = \lambda_I \widetilde{I}_z(x(u))$, then the KL-regularized generator objective is exactly

$$\pi \mapsto \lambda_I \widetilde{\mathcal{D}}_z(\pi) - \mathcal{C}_{L,z}(\pi). \tag{42}$$

A maximizer of (42) exists because the finite-horizon policy space is a finite product of compact simplices and the objective is continuous. Let $\pi_z^\dagger$ be any such maximizer. Then for every policy π,

$$\lambda_I \widetilde{\mathcal{D}}_z(\pi_z^\dagger) - \mathcal{C}_{L,z}(\pi_z^\dagger) \ge \lambda_I \widetilde{\mathcal{D}}_z(\pi) - \mathcal{C}_{L,z}(\pi). \tag{43}$$

By Assumption 4.7, for every policy ν,

$$\left|\widetilde{\mathcal{D}}_z(\nu) - \mathcal{D}_z(\nu)\right| \le \mathcal{E}_z(\nu), \tag{44}$$

because $\Gamma_L(U_\nu) \in [0,1]$ and the pointwise error is bounded by $\delta_z(x(U_\nu))$. Using (44) with $\nu = \pi_z^\dagger$ gives

$$\lambda_I \mathcal{D}_z(\pi_z^\dagger) - \mathcal{C}_{L,z}(\pi_z^\dagger) \geq \lambda_I \widetilde{\mathcal{D}}_z(\pi_z^\dagger) - \mathcal{C}_{L,z}(\pi_z^\dagger) - \lambda_I \mathcal{E}_z(\pi_z^\dagger).$$

Combining this with (43) and then applying (44) to $\nu = \pi$ yields

$$\begin{aligned}\lambda_I \mathcal{D}_z(\pi_z^\dagger) - \mathcal{C}_{L,z}(\pi_z^\dagger) &\geq \lambda_I \widetilde{\mathcal{D}}_z(\pi) - \mathcal{C}_{L,z}(\pi) - \lambda_I \mathcal{E}_z(\pi_z^\dagger) \\ &\geq \lambda_I \mathcal{D}_z(\pi) - \mathcal{C}_{L,z}(\pi) - \lambda_I \big(\mathcal{E}_z(\pi) + \mathcal{E}_z(\pi_z^\dagger)\big),\end{aligned}$$

which proves (12). If $\delta_z \equiv 0$, then $\mathcal{E}_z(\pi) = 0$ for all π, and the displayed inequality becomes exact optimality of expected entropy reduction minus KL cost.

Part (c): composite reward objective. For the composite reward, the KL-regularized generator objective is

$$\pi \mapsto \mathcal{O}_z(\pi) + \lambda_I \widetilde{\mathcal{D}}_z(\pi) - \mathcal{C}_{L,z}(\pi). \tag{45}$$

Again, compactness and continuity imply that a maximizer exists. Let $\pi_z^\star$ maximize (45). Then, for every π,

$$\mathcal{O}_z(\pi_z^\star) + \lambda_I \widetilde{\mathcal{D}}_z(\pi_z^\star) - \mathcal{C}_{L,z}(\pi_z^\star) \geq \mathcal{O}_z(\pi) + \lambda_I \widetilde{\mathcal{D}}_z(\pi) - \mathcal{C}_{L,z}(\pi). \tag{46}$$

Applying the same surrogate-error bound (44) to $\pi_z^\star$ and to π gives exactly the stated composite robust inequality. This also proves the final interpretation: the IG component has a precise entropy-reduction meaning by part (a), while the additional reward terms modify the optimized utility and therefore prevent any claim of pure information-gain optimality in the composite setting.

Part (d): the IG-ablated comparator. Fix $\lambda_I > 0$. The objective $\mathcal{O}_z - \mathcal{C}_{L,z}$ is continuous on the finite product of policy simplices: trajectory probabilities are polynomials in policy coordinates and $p\log(p/\rho)$ extends continuously at $p = 0$ when $\rho > 0$. It attains its maximum. Choose any maximizer π_z^0; no uniqueness assumption is needed. By its optimality, $\mathfrak{G}_z \geq 0$. Composite optimality against this comparator yields

$$\lambda_I \big[\widetilde{\mathcal{D}}_z(\pi_z^\star) - \widetilde{\mathcal{D}}_z(\pi_z^0)\big] \geq (\mathcal{O}_z - \mathcal{C}_{L,z})(\pi_z^0) - (\mathcal{O}_z - \mathcal{C}_{L,z})(\pi_z^\star) = \mathfrak{G}_z.$$

Apply (44) to each policy and divide by the strictly positive λ_I to obtain (14). The comparison removes the explicit IG term only; frontier and novelty can remain in $\mathcal{O}_z$. This is why π_z^0 can depend on z. □

A.8 Proof of Corollary A.3

Corollary A.3 (Coarse-grained terminal Gibbs weights)**.** *For a finite terminal-law relaxation with constant $\alpha > 0$, full-support ν_{ref}, and reward $\bar{R}(\chi(u))$ constant on each nonempty partition cell, maximizing $\mathbb{E}_\nu \bar{R} - \alpha \mathrm{KL}(\nu \| \nu_{\mathrm{ref}})$ gives*

$$\mu^*(c) = \frac{\mu_{\mathrm{ref}}(c) e^{\bar{R}(c)/\alpha}}{\sum_{c'} \mu_{\mathrm{ref}}(c') e^{\bar{R}(c')/\alpha}}, \tag{47}$$

and leaves the reference conditional within each cell unchanged.

Proof. The proof is a finite-dimensional Gibbs variational calculation, but we give the details to make clear what is and is not implied. Define

$$Z := \sum_{\omega \in \Omega_L} \nu_{\mathrm{ref}}(\omega) \exp(\bar{R}_z(\chi(\omega))/\alpha).$$

Since Ω_L is finite, ν_{ref} has full support, and $\bar{R}_z$ is bounded, $0 < Z < \infty$. Let ν_z^* be the terminal Gibbs distribution

$$\nu_z^*(\omega) = \frac{\nu_{\mathrm{ref}}(\omega) \exp(\bar{R}_z(\chi(\omega))/\alpha)}{\sum_{\omega' \in \Omega_L} \nu_{\mathrm{ref}}(\omega') \exp(\bar{R}_z(\chi(\omega'))/\alpha)}. \tag{48}$$

For any $\nu \in \Delta(\Omega_L)$,

$$\begin{aligned}\sum_{\omega} \nu(\omega)\bar{R}_z(\chi(\omega)) - \alpha\mathrm{KL}(\nu\|\nu_{\mathrm{ref}}) &= \sum_{\omega} \nu(\omega)\bar{R}_z(\chi(\omega)) - \alpha\sum_{\omega} \nu(\omega)\log\frac{\nu(\omega)}{\nu_{\mathrm{ref}}(\omega)} \\ &= \alpha\log Z - \alpha\sum_{\omega} \nu(\omega)\log\frac{\nu(\omega)}{\nu_z^*(\omega)} \\ &= \alpha\log Z - \alpha\mathrm{KL}(\nu\|\nu_z^*).\end{aligned}$$

Here terms with $\nu(\omega) = 0$ are interpreted using the convention $0\log 0 = 0$. Since $\mathrm{KL}(\nu\|\nu_z^*) \geq 0$, equality is attained uniquely at $\nu = \nu_z^*$. This proves uniqueness and (48).

Summing (48) over all outcomes in cell c yields

$$\begin{aligned}\mu_z^*(c) &= \sum_{\omega:\chi(\omega)=c} \frac{\nu_{\mathrm{ref}}(\omega)\exp(\bar{R}_z(c)/\alpha)}{Z} \\ &= \frac{\mu_{\mathrm{ref}}(c)\exp(\bar{R}_z(c)/\alpha)}{Z}.\end{aligned}$$

Rewriting Z by grouping outcomes into cells gives

$$Z = \sum_{c'\in\mathcal{C}} \mu_{\mathrm{ref}}(c')\exp(\bar{R}_z(c')/\alpha),$$

which proves (47). The conditional distribution within a cell is

$$\nu_z^*(\omega \mid \chi(\omega) = c) = \frac{\nu_z^*(\omega)}{\mu_z^*(c)} = \frac{\nu_{\mathrm{ref}}(\omega)}{\mu_{\mathrm{ref}}(c)} = \nu_{\mathrm{ref}}(\omega \mid \chi(\omega) = c). \tag{49}$$

which proves the reference-conditional assertion.

The cell-level equivalence can be proved either by repeating the same Gibbs calculation on the finite simplex $\Delta(\mathcal{C})$ or by decomposing KL. For completeness, we show the decomposition. Any distribution $\nu \in \Delta(\Omega_L)$ with cell marginal μ satisfies

$$\mathrm{KL}(\nu\|\nu_{\mathrm{ref}}) = \mathrm{KL}(\mu\|\mu_{\mathrm{ref}}) + \sum_{c\in\mathcal{C}} \mu(c)\mathrm{KL}\big(\nu(\cdot \mid c)\|\nu_{\mathrm{ref}}(\cdot \mid c)\big), \tag{50}$$

with the convention that the conditional term is zero when $\mu(c) = 0$. Since the reward depends only on c, for a fixed cell marginal μ the best within-cell conditional is the reference conditional, making the second term in (50) vanish. Therefore the outcome-level problem reduces exactly to the cell-level relaxation, whose unique solution is (47). Finally, the log-odds relation follows by dividing (47) for cells c and c' and taking logarithms. □

A.9 Proof of Lemma A.4

Lemma A.4 (Frozen fast stability)**.** *For fixed L, z, the ODE*

$$\dot{q} = \mathcal{T}_z q - q \tag{51}$$

has unique global solutions and equilibrium $q^*(z)$, with

$$\|q(t) - q^*(z)\|_\infty \leq e^{-(1-\gamma)t}\|q(0) - q^*(z)\|_\infty. \tag{52}$$

Under Assumption 4.9, $\mathcal{Q}_L$ is forward invariant.

Proof. The proof is fixed-horizon throughout. Thus $\mathcal{X}_L$ is finite and $\mathbb{R}^{\mathcal{X}_L}$ equipped with $\|\cdot\|_\infty$ is a finite-dimensional Banach space.

First, $\mathcal{T}_z$ is a γ-contraction in sup norm by Proposition 4.1: for all $q, q' \in \mathbb{R}^{\mathcal{X}_L}$,

$$\|\mathcal{T}_z q - \mathcal{T}_z q'\|_\infty \leq \gamma\|q - q'\|_\infty. \tag{53}$$

Consequently the vector field $F_z(q) = \mathcal{T}_z q - q$ is globally Lipschitz:

$$\|F_z(q) - F_z(q')\|_\infty \leq (1+\gamma)\|q - q'\|_\infty.$$

Hence the ODE has a unique solution for every initial condition, and the solution cannot explode in finite time because the vector field has at most linear growth. An equilibrium of (51) is precisely a fixed point of $\mathcal{T}_z$. By the contraction mapping theorem this fixed point is unique, namely $q^*(z)$.

It remains to prove the exponential estimate. Let

$$e(t) := q(t) - q^*(z), \qquad V(t) := \|e(t)\|_\infty.$$

Since $q^*(z) = \mathcal{T}_z q^*(z)$, for any $h > 0$,

$$\begin{aligned} e(t) + h\dot{e}(t) &= e(t) + h\big(\mathcal{T}_z q(t) - q(t)\big) \\ &= (1-h)e(t) + h\big(\mathcal{T}_z q(t) - \mathcal{T}_z q^*(z)\big). \end{aligned}$$

For $0 < h \le 1$, using (53),

$$\begin{aligned} \|e(t) + h\dot{e}(t)\|_\infty &\le (1-h)\|e(t)\|_\infty + h\|\mathcal{T}_z q(t) - \mathcal{T}_z q^*(z)\|_\infty \\ &\le \big(1 - (1-\gamma)h\big)\|e(t)\|_\infty. \end{aligned}$$

Because q is differentiable, $e(t+h) = e(t) + h\dot{e}(t) + o(h)$ in sup norm. Therefore the upper right Dini derivative of V satisfies

$$D^+V(t) := \limsup_{h \downarrow 0} \frac{V(t+h) - V(t)}{h} \le -(1-\gamma)V(t).$$

Because F_z is Lipschitz, q is continuously differentiable and V is locally Lipschitz. Hence V is differentiable almost everywhere. At every differentiability point the above Dini inequality gives

$$V'(t) \le -(1-\gamma)V(t).$$

Therefore $t \mapsto e^{(1-\gamma)t}V(t)$ has nonpositive derivative almost everywhere, and so it is nonincreasing. Consequently

$$V(t) \le e^{-(1-\gamma)t}V(0),$$

which is exactly (52). Hence $q^*(z)$ is globally exponentially stable, and in particular globally asymptotically stable.

Finally we verify forward invariance of the stability box. If $q \in \mathcal{Q}_L$, then for every coordinate (s,a),

$$|(\mathcal{T}_z q)(s,a)| \le R_{\max} + \gamma\|q\|_\infty \le R_{\max} + \gamma C_Q \le C_Q,$$

where the last inequality uses $C_Q \ge R_{\max}/(1-\gamma)$. At an upper boundary coordinate $q(s,a) = C_Q$, this implies

$$\dot{q}(s,a) = (\mathcal{T}_z q)(s,a) - C_Q \le 0,$$

and at a lower boundary coordinate $q(s,a) = -C_Q$ it implies

$$\dot{q}(s,a) = (\mathcal{T}_z q)(s,a) + C_Q \ge 0.$$

Thus the vector field points inward on every face of the box, and $\mathcal{Q}_L$ is forward invariant. □

A.10 Proof of Lemma A.5

Lemma A.5 (Lipschitz equilibrium map)**.** *Under Assumption 4.9, with* $C_T = L_r + 2\gamma C_Q L_\alpha/\alpha_{\min}$,

$$\|q^*(z) - q^*(z')\|_\infty \le \frac{C_T}{1-\gamma}\|z - z'\|. \tag{54}$$

Proof. The proof is again fixed-horizon. Thus $\mathcal{X}_L$ is finite, $q^*(z)$ is a finite table, and all norms on the slow state space are equivalent up to constants. We use the Euclidean norm on $\mathcal{Z}$ and the sup norm on $\mathbb{R}^{\mathcal{X}_L}$.

Step 1: Lipschitz dependence of the soft value on the temperature. Fix a valid nonterminal prefix s and a bounded vector $q \in [-C_Q, C_Q]^{\mathcal{A}(s)}$. For $\alpha \in [\alpha_{\min}, \alpha_{\max}]$, define

$$W_s(q, \alpha) := \alpha \log \sum_{a \in \mathcal{A}(s)} \pi_{\mathrm{ref}}(a \mid s) \exp(q(a)/\alpha). \tag{55}$$

We claim that

$$|W_s(q, \alpha) - W_s(q, \alpha')| \le \frac{2C_Q}{\alpha_{\min}} |\alpha - \alpha'|. \tag{56}$$

To see this, write

$$Z(\alpha) := \sum_a \pi_{\mathrm{ref}}(a \mid s) e^{q(a)/\alpha}, \qquad g_\alpha(a) := \frac{\pi_{\mathrm{ref}}(a \mid s) e^{q(a)/\alpha}}{Z(\alpha)}.$$

The function $\alpha \mapsto W_s(q, \alpha)$ is continuously differentiable on $[\alpha_{\min}, \alpha_{\max}]$ and

$$\frac{\partial}{\partial \alpha} W_s(q, \alpha) = \log Z(\alpha) - \frac{1}{\alpha} \sum_a g_\alpha(a) q(a).$$

Because $q(a) \in [-C_Q, C_Q]$ and $\pi_{\mathrm{ref}}(\cdot \mid s)$ is a probability distribution,

$$e^{-C_Q/\alpha} \le Z(\alpha) \le e^{C_Q/\alpha}, \qquad \left| \sum_a g_\alpha(a) q(a) \right| \le C_Q.$$

Therefore

$$\left| \frac{\partial}{\partial \alpha} W_s(q, \alpha) \right| \le \frac{2C_Q}{\alpha} \le \frac{2C_Q}{\alpha_{\min}},$$

and the mean value theorem proves (56). If the successor state is terminal, the continuation value is identically zero and the same bound holds trivially.

Step 2: Lipschitz dependence of the Bellman operator on the slow state. Let $q \in \mathcal{Q}_L$. For a coordinate $(s, a) \in \mathcal{X}_L$, using the definition of the soft Bellman operator,

$$\begin{aligned} |(\mathcal{T}_z q)(s, a) - (\mathcal{T}_{z'} q)(s, a)| \le{}& |r_z(s, a) - r_{z'}(s, a)| \\ &+ \gamma \, |V_{z,q}(s \oplus a) - V_{z',q}(s \oplus a)|, \end{aligned}$$

where $V_{z,q}$ denotes the soft value computed with temperature α_z. The reward term is bounded by the Lipschitz reward assumption. If $s \oplus a$ is terminal, the value term is zero. Otherwise, applying (56) and the Lipschitz temperature assumption,

$$|V_{z,q}(s \oplus a) - V_{z',q}(s \oplus a)| \le \frac{2C_Q}{\alpha_{\min}} L_\alpha \|z - z'\|.$$

Define

$$C_T := L_r + \gamma \frac{2C_Q}{\alpha_{\min}} L_\alpha. \tag{57}$$

Taking the supremum over (s, a) gives the uniform operator bound

$$\|\mathcal{T}_z q - \mathcal{T}_{z'} q\|_\infty \le C_T \|z - z'\|, \qquad q \in \mathcal{Q}_L. \tag{58}$$

Step 3: Lipschitz continuity of the fixed point. By Proposition 4.1, for every $z \in \mathcal{Z}$,

$$\|q^*(z)\|_\infty \le \frac{R_{\max}}{1 - \gamma} \le C_Q,$$

so $q^*(z) \in \mathcal{Q}_L$. For any $z, z' \in \mathcal{Z}$,

$$\begin{aligned} \|q^*(z) - q^*(z')\|_\infty &= \|\mathcal{T}_z q^*(z) - \mathcal{T}_{z'} q^*(z')\|_\infty \\ &\le \|\mathcal{T}_z q^*(z) - \mathcal{T}_z q^*(z')\|_\infty + \|\mathcal{T}_z q^*(z') - \mathcal{T}_{z'} q^*(z')\|_\infty \\ &\le \gamma \|q^*(z) - q^*(z')\|_\infty + C_T \|z - z'\|, \end{aligned}$$

where the last line uses the contraction of $\mathcal{T}_z$ in its q argument and (58) with $q = q^*(z') \in \mathcal{Q}_L$. Rearranging yields

$$(1 - \gamma) \|q^*(z) - q^*(z')\|_\infty \le C_T \|z - z'\|,$$

which proves (54). □

A.11 REALIZED-ROW SUPPORT AND THE COUPLED CERTIFICATE

We first prove the sharper support bound, then the residual-to-law estimate, and finally the information transfer. The support proof does not use proximity to a Bellman fixed point.

Lemma A.6 (Realized-row completed-path floor)**.** *Let q be any finite table on the active prefix tree, $\alpha(s) > 0$, and let $\pi_{q,z}$ use a full-support reference and the same legal mask. For every completed path $u = (a_0, \dots, a_{\ell(u)-1})$,*

$$\mu_{q,z}(u) \geq \pi_{\mathrm{ref}}(u) \exp\left[- \sum_{k<\ell(u)} \frac{\max_{b\in\mathcal{A}(s_k)} q(s_k, b) - q(s_k, a_k)}{\alpha(s_k)} \right]. \tag{59}$$

If $\|q\|_\infty \leq C_Q$ and $\alpha(s) \geq \alpha_{\min}$, the right-hand side is at least $\pi_{\mathrm{ref}}(u) e^{-2C_Q \ell(u)/\alpha_{\min}}$.

Proof. For a fixed row s, let $m_s = \max_b q(s, b)$. Since the legal reference probabilities sum to one,

$$Z_s = \sum_b \pi_{\mathrm{ref}}(b \mid s) e^{q(s,b)/\alpha(s)} \leq e^{m_s/\alpha(s)}.$$

It follows for each legal a that

$$\pi_{q,z}(a \mid s) = \pi_{\mathrm{ref}}(a \mid s) e^{q(s,a)/\alpha(s)} / Z_s \geq \pi_{\mathrm{ref}}(a \mid s) e^{-(m_s - q(s,a))/\alpha(s)}.$$

Multiplication along the autoregressive path gives (59), including its EOS action. A deterministic padding row contributes a ratio of one and gap zero. Under the box condition, every $m_s - q(s, a)$ lies in $[0, 2C_Q]$, which proves the second assertion. The inequalities need not be strict; equal action values are an example of equality. □

Current versus future support. For adapted q_t, z_t, the preceding bound is a pre-generation conditional bound. On a realized finite record, $1/\kappa_t^{\mathrm{row}}(u)$ is not generally an upper bound on the remaining waiting time under future policy updates. If the current policy is frozen, independent resampling has mean waiting time $1/\mu_t(u) \leq 1/\kappa_t^{\mathrm{row}}(u)$. With continued adaptation one instead needs a lower bound valid over the future times considered. The time-uniform box floor provides precisely this in Theorem 4.4. Bounded semantic classes inherit the same argument by selecting a member, or by summing the floors of finitely many distinct eligible members, whose one-round events are disjoint. No unrestricted semantic equivalence test is used.

Lemma A.7 (Gibbs and path-law stability)**.** *Fix a finite prefix tree of depth at most L, common reference, mask, and temperature function with $\alpha \geq \alpha_{\min}$. For two finite tables q, q', let $d = \|q - q'\|_\infty$. Their full terminal-outcome laws satisfy*

$$\left| \log \frac{\pi_{q,z}(a \mid s)}{\pi_{q',z}(a \mid s)} \right| \leq \frac{2d}{\alpha_{\min}}, \qquad \mathrm{TV}(\mu_{q,z}, \mu_{q',z}) \leq \tanh(Ld/\alpha_{\min}). \tag{60}$$

Proof. At a common row of temperature α_s, the inequalities

$$e^{-d/\alpha_s} Z_s(q') \leq Z_s(q) \leq e^{d/\alpha_s} Z_s(q')$$

follow termwise. Subtracting the two log normalizers in the log-policy ratio therefore gives absolute value at most $2d/\alpha_s$. For each distinguishable terminal path $\omega \in \Omega_L$, multiply the action probabilities. Its length is at most L, hence

$$e^{-c} \leq \frac{\mu_{q,z}(\omega)}{\mu_{q',z}(\omega)} \leq e^c, \qquad c = 2Ld/\alpha_{\min}.$$

Both are normalized distributions on the same finite space, with the same support. For completeness, let $P/Q = R$, $a = e^{-c}$ and $b = e^c$. Since $\mathbb{E}_Q R = 1$ and the convex function $(R-1)_+$ lies below its chord between a and b,

$$\mathrm{TV}(P, Q) = \mathbb{E}_Q (R-1)_+ \leq \frac{(b-1)(1-a)}{b-a} = \frac{e^c - 1}{e^c + 1} = \tanh(c/2).$$

The case $c = 0$ is equality of the laws. If terminal paths are mapped to a coarser outcome, TV cannot increase under the mapping, so the same bound remains valid. Conditioning on completion would instead change the distributions and is deliberately not done. □

Proof of Theorem 4.12. Part (i). Apply Lemma A.6 at the current q_t, z_t. The bound is unconditional on future optimizer performance and is valid even if the projected box is too small to contain the ideal fixed point. The larger invariant-box condition in Assumption 4.9 is needed only for the subsequent asymptotic tracking theorem.

Part (ii). Since $q^* = \mathcal{T}_{z_t} q^*$,

$$\|q_t - q^*\|_\infty \le \|q_t - \mathcal{T}_{z_t} q_t\|_\infty + \|\mathcal{T}_{z_t} q_t - \mathcal{T}_{z_t} q^*\|_\infty \le r_t + \gamma \|q_t - q^*\|_\infty.$$

Thus $d_t := \|q_t - q^*\|_\infty \le r_t/(1-\gamma)$. Substitute this into Lemma A.7, which uses the same reference, mask, and temperature at both policies, to prove (18).

Part (iii). Proposition 4.8 gives (14). The function $f_t(\omega) = \Gamma_L(\omega) I_{z_t}(x(\omega))$ lies in $[0, I_{\max}]$. For any two probability laws P, Q and any $0 \le f \le M$,

$$|\mathbb{E}_P f - \mathbb{E}_Q f| \le M \mathrm{TV}(P, Q).$$

One proof divides by M and uses $\mathbb{E}f/M = \int_0^1 P(f/M > s)\, ds$; the zero-$M$ case is immediate. Therefore

$$\mathcal{D}_{z_t}(\pi_t) \ge \mathcal{D}_{z_t}(\pi_t^\star) - I_{\max} \varepsilon_t^{\mathrm{tr}}.$$

Combining with (14) proves (19). As its TV coefficient is negative, replacing $\varepsilon_t^{\mathrm{tr}}$ by its upper bound gives a (possibly looser) valid lower bound. In general $\mathfrak{G}_z$ is an objective gap, not a terminal KL divergence; only a separately justified terminal-law specialization identifies the two. □

A.12 Sampled Bellman residuals and confidence scope

The deterministic certificate uses the mean operator. This subsection states explicit sampling conditions under which an empirical residual can be substituted; they are not guaranteed by a generic neural TD implementation.

Lemma A.8 (Finite-coordinate empirical residual)**.** *Condition on a fixed pre-validation state (q, z) with $\|q\|_\infty \le C_Q$. Let $d_L = |\mathcal{X}_L|$. Suppose that for every coordinate i one can obtain n conditionally independent Bellman-target samples $X_{i,j} \in [-M, M]$ with $\mathbb{E}[X_{i,j} \mid q, z] = (\mathcal{T}_z q)_i$. Coordinates need not be mutually independent. For $\widehat{\mathcal{T}}_z q$ formed from coordinatewise means and $0 < \rho < 1$, set*

$$c = M\sqrt{\frac{2\log(2d_L/\rho)}{n}}, \qquad \widehat{r} = \|q - \widehat{\mathcal{T}}_z q\|_\infty.$$

With conditional probability at least $1 - \rho$,

$$\|\widehat{\mathcal{T}}_z q - \mathcal{T}_z q\|_\infty \le c, \qquad r \le \widehat{r} + c, \qquad \mathrm{TV}(\mu_{q,z}, \mu_{q^*,z}) \le \tanh\!\left(\frac{L(\widehat{r}+c)}{\alpha_{\min}(1-\gamma)}\right). \tag{61}$$

Proof. For each coordinate, the bounded-mean exponential inequality gives

$$\mathbb{P}\left(\left|n^{-1}\sum_j X_{i,j} - \mathbb{E}X_{i,1}\right| > c \mid q, z\right) \le 2\exp\{-nc^2/(2M^2)\} = \rho/d_L.$$

A union bound yields the first assertion; the triangle inequality yields the second. Part (ii) of Theorem 4.12 gives the last assertion on the same event. For rewards bounded by $R_{\max}$ and bootstrap values in $[-C_Q, C_Q]$, $M = R_{\max} + \gamma C_Q$ is sufficient. On a deterministic prefix model the exact operator may instead be computed, giving $c = 0$. □

Validation samples in this lemma are drawn after q, z are fixed. Repeatedly choosing q using the same validation samples requires a uniform or sequential bound not supplied by the lemma. Across rounds or validation calls, conditional error budgets ρ_t with $\sum_t \rho_t \le \rho_{\mathrm{total}}$ give a simultaneous event by a union bound. A confidence miss is recorded, not removed from the dataset. Estimator confidence and Bellman-residual confidence are separate channels and must be budgeted separately if used jointly.

A.13 Proof of the finite Bayesian entropy budget

Proof of Corollary 4.13. Fix an initial belief w_1 and component observation models. A sufficient coherent construction draws a latent M from w_1 and then, conditionally on M and chosen queries, samples observations from the specified likelihoods. The generator and any model-based estimator simulations use only the observed history and independent randomization, not additional information about M. The stated predictive law then follows from Bayes' rule. The corollary also applies to any process directly satisfying that predictive-law hypothesis.

For a fixed current outcome u, Lemma 4.6 gives

$$\mathbb{E}[H(w_t) - H(w_{t+1}) \mid \mathcal{F}_{t-1}, U_t = u] = I_{z_t}(x(u)) \geq 0.$$

Taking expectation over the chosen program produces the unweighted expected drop $d_t = \mathbb{E}_{\mu_t} I_{z_t}(x(U_t))$. Because $0 \leq \Gamma_L \leq 1$,

$$d_t \geq \mathbb{E}_{\mu_t}[\Gamma_L(U_t) I_{z_t}(x(U_t))] = \mathcal{D}_{z_t}(\pi_t) \geq A_t - e_t.$$

Taking outer expectations and summing from 1 to T telescopes the left-hand side, yielding (20). Entropy is bounded by $\log K$, and the stated integrability controls the error terms. The same proof applies with B_t^{TV} in place of $\varepsilon_t^{\mathrm{tr}}$, at the cost of the larger tracking penalty.

A probabilistic estimator envelope establishes this inequality on its current coverage event, not unconditionally by changing the sampling law after the event is observed. For an unconditional expected budget with a deterministic conditional failure bound ρ_t, one safe alternative is to use the universal clipped error $I_{\max}$ on failed events and include a corresponding failure penalty (at most $2I_{\max}\rho_t$ for the two surrogate expectations). This requires the current-state event bound to hold conditionally before the estimator draw. The deterministic-envelope corollary itself does not hide this extra term. □

Trajectory audit. For a realized trajectory, $H(w_1) - H(w_{T+1}) = \sum_t [H(w_t) - H(w_{t+1})]$ is an exact telescope, but the realized increments need not equal d_t. Complete enumeration of the observation tree verifies the expected identity exactly up to numerical precision. Monte Carlo trajectories instead estimate its expectation, with sampling uncertainty. Fixed-target benchmark runs with unbalanced target frequencies are not exact prior-predictive draws and therefore are not a numerical proof of the unconditional corollary; they can still audit each model-based one-step predictive quantity and compare learning empirically.

A.14 Projected interpolation and proof of Theorem 4.15

We give the projection argument explicitly because the projected drift need not be continuous at the boundary. Throughout, L is fixed; all norms on the finite spaces are equivalent, with constants independent of time.

Lemma A.9 (Projected interpolation on a compact convex set)**.** *Let $K \subset \mathbb{R}^d$ be compact and convex and let $f : K \to \mathbb{R}^d$ be Lipschitz. Consider*

$$x_{n+1} = \Pi_K\{x_n + c_n[f(x_n) + \zeta_{n+1} + e_n]\},$$

where $c_n \to 0$, $\sum_n c_n = \infty$, ζ_{n+1} is bounded, $e_n \to 0$, and the cumulative sums $\sum_{k<n} c_k \zeta_{k+1}$ converge. Then its piecewise-linear interpolation is an asymptotic pseudo-trajectory of the unique continuous semiflow solving

$$\dot{x} = f(x) - v, \qquad v \in N_K(x), \qquad x \in K,$$

where $N_K(x) = \{v : \langle v, y - x\rangle \leq 0 \text{ for all } y \in K\}$. Almost everywhere, this is $\dot{x} = \Pi_{T_K(x)} f(x)$. The same finite-window conclusion holds for a uniformly Lipschitz compact family f_ϑ when its parameter varies by $o(1)$ on each such window.

Proof. Set $F_n = f(x_n) + \zeta_{n+1} + e_n$ and

$$v_n = F_n - (x_{n+1} - x_n)/c_n.$$

The Euclidean projection optimality condition gives $v_n \in N_K(x_{n+1})$. Since $x_n \in K$, $\|x_{n+1} - x_n\| \leq c_n\|F_n\|$ and $\|v_n\| = \operatorname{dist}(x_n + c_n F_n, K)/c_n \leq \|F_n\|$. All these bounds are uniform on a tail because f and the noise are bounded and $e_n \to 0$.

Let $s_n = \sum_{k<n} c_k$. On any shifted interval $[s_n, s_n + T]$, both the interpolation of x and the integral of the piecewise-constant v_k are uniformly Lipschitz. Select a subsequence on which the initial points converge. Compactness and Arzelà–Ascoli give a further subsequence on which these two paths converge uniformly to $x(\cdot)$ and an absolutely continuous correction path. The derivatives of the latter converge weakly along a further subsequence to a bounded $v(\cdot)$. The noise integrals vanish uniformly: convergence of the cumulative noise series controls full increments, and at most two incomplete mesh intervals contribute $O(\sup_{k\geq n} c_k)$. Similarly the integrals of e_k vanish. Lipschitz continuity of f and vanishing mesh size therefore imply

$$x(t) = x(0) + \int_0^t [f(x(s)) - v(s)]\, ds.$$

For every $y \in K$, the inequality $\langle v_k, y - x_{k+1}\rangle \leq 0$ holds on each mesh interval. Integrate it against a nonnegative continuous test function. Uniform convergence of x and weak convergence of the bounded v_k pass it to the limit. A countable dense subset of K then yields $v(s) \in N_K(x(s))$ for almost every s.

At a two-sided differentiability point of the feasible path x, testing the normal inequality with $x(s + h)$ for both signs of h gives $\langle v(s), \dot{x}(s)\rangle = 0$. Also $\dot{x}(s) \in T_K(x(s))$. The normal/tangent polar decomposition consequently gives $\dot{x}(s) = \Pi_{T_K(x(s))} f(x(s))$ almost everywhere. Existence for every initial point follows by the same compactness argument applied to deterministic projected Euler approximations.

For two solutions (x, v) and (y, w), monotonicity of the normal cone gives $\langle x - y, v - w\rangle \geq 0$. Hence, with a Lipschitz constant L_f,

$$\frac{d}{dt}\|x - y\|^2 \leq 2\langle x - y, f(x) - f(y)\rangle \leq 2L_f\|x - y\|^2$$

almost everywhere. Gronwall proves uniqueness and continuous dependence on initial points, and hence a continuous semiflow. If uniform finite-window tracking failed, the compactness construction on a violating sequence would give a limiting solution different from this unique solution with the same initial point, a contradiction. Arbitrary interpolation start times differ from mesh start times by a vanishing interval.

For the parameterized statement, take a convergent subsequence of the parameters at window starts. Uniform parameter continuity and the stipulated $o(1)$ variation replace $f_{\vartheta_k}(x_k)$ by its frozen counterpart in the integral with an $o(1)$ error. The same argument and uniqueness apply. □

Proof of Theorem 4.15. **Step 1: the averaged drift and noise tails.** The equilibrium regularity gives

$$\|\bar{h}(z) - \bar{h}(z')\| \leq L_h \left(1 + \frac{C_T}{1 - \gamma}\right) \|z - z'\|. \tag{62}$$

Thus $\bar{h}$ is Lipschitz and bounded. With the filtration of Assumption 4.10, the sums

$$S_n^q = \sum_{k=1}^{n-1} a_k M_{k+1}^q, \qquad S_n^z = \sum_{k=1}^{n-1} b_k M_{k+1}^z$$

are martingales relative to $\mathcal{F}_{n-1}$. Bounded conditional second moments and square-summable steps give bounded L^2 martingales coordinatewise, hence almost-sure and L^2 convergence. Consequently, for every $T < \infty$,

$$\sup_{m\geq n:\, \sum_{k=n}^{m-1} a_k \leq T} \left\| \sum_{k=n}^{m-1} a_k M_{k+1}^q \right\|_\infty \longrightarrow 0, \tag{63}$$

$$\sup_{m\geq n:\, \sum_{k=n}^{m-1} b_k \leq T} \left\| \sum_{k=n}^{m-1} b_k M_{k+1}^z \right\| \longrightarrow 0. \tag{64}$$

These limits hold simultaneously for rational T, and then for all finite T by monotonicity.

Step 2: slow-state variation on a fast window. Projection nonexpansiveness and bounded drifts/noise give, on an almost-sure tail, $\|z_{k+1} - z_k\| \le C b_k$. If the fast elapsed time from n to m is at most T plus one mesh step, then

$$\sup_m \|z_m - z_n\| \le C(T + o(1)) \sup_{k \ge n} \frac{b_k}{a_k} \longrightarrow 0.$$

In particular,

$$\sup_{n \le m:\, \sum_{k=n}^{m-1} a_k \le T} \|z_m - z_n\| \to 0. \tag{65}$$

The random bound on ϵ_k on a tail is finite because $\epsilon_k \to 0$ a.s.

Step 3: projected fast tracking of the frozen flow. Let $s_n^a = \sum_{k=1}^{n-1} a_k$ and $\bar{q}^a(s_n^a) = q_n$. The vector fields $F_z(q) = \mathcal{T}_z q - q$ are uniformly Lipschitz in q and, on the invariant box, uniformly Lipschitz in z by (58). Apply Lemma A.9 using (63) and (65). By Lemma A.4, $F_z(q)$ is inward-pointing on the box, so its projected flow coincides with its unprojected flow Φ^z. We obtain

$$\sup_{0 \le u \le T} \|\bar{q}^a(s_n^a + u) - \Phi_u^{z_n}(q_n)\|_\infty \to 0 \quad \text{a.s.} \tag{66}$$

This step does not interchange projection and conditional expectation; the normal-cone proof handles boundary noise before passing to the limit.

Step 4: convergence to the moving equilibrium. Let D_Q be the box diameter in sup norm. For arbitrary $\varepsilon > 0$, choose a fixed T with $e^{-(1-\gamma)T} D_Q < \varepsilon/3$. For each sufficiently large n, choose a past mesh index $k(n)$ for which $s_n^a - s_{k(n)}^a = T + o(1)$. Such indices exist, tend to infinity, and have vanishing overshoot because $a_n \to 0$. By (66),

$$\|q_n - q^*(z_{k(n)})\|_\infty \le e^{-(1-\gamma)(T+o(1))} D_Q + o(1) < 2\varepsilon/3.$$

Equilibrium Lipschitzness and (65) give $\|q^*(z_n) - q^*(z_{k(n)})\|_\infty = o(1)$. Thus the error is below ε for every sufficiently large n, not only selected block endpoints. This proves (22).

Step 5: the slow projected semiflow. By tracking,

$$\|h(q_t, z_t) - \bar{h}(z_t)\| \le L_h \|q_t - q^*(z_t)\|_\infty \to 0. \tag{67}$$

This discrepancy plus ϵ_t is a vanishing bias. Apply Lemma A.9 to the slow recursion with drift $\bar{h}$ and noise tails (64). Its unique continuous semiflow is denoted Ψ. If $\bar{z}$ is the slow interpolation, then

$$\lim_{s \to \infty} \sup_{0 \le u \le T} \|\bar{z}(s + u) - \Psi_u(\bar{z}(s))\| = 0 \quad (T < \infty). \tag{68}$$

Thus the limit is an APT of a single well-defined projected semiflow, not an unspecified set-valued approximation.

Step 6: limit sets and the Lyapunov qualification. Compactness makes the interpolation precompact. The APT limit-set theorem for a continuous semiflow gives a nonempty compact invariant internally chain transitive limit set (Benaïm, 1999, Theorem 5.7). If V is strict relative to the stationary set $\mathcal{S}$ and $V(\mathcal{S})$ has empty interior, the chain-transitivity Lyapunov criterion places the limit set in $\mathcal{S}$ (Benaïm, 1999, Proposition 6.4). Equivalently, $\operatorname{dist}(z_t, \mathcal{S}) \to 0$. The empty-interior qualification cannot in general be dropped for an arbitrary continuous Lyapunov function. □

A.15 Proof of Theorem 4.16

Proof. We prove the three parts separately.

Part (a): existence of finite tracking epochs. Fix an epoch e and condition on the past at its start. Since $N_e < \infty$ and $q_{N_e}^{(e)} \in \mathcal{Q}_{L_e}$, $z_{N_e} \in \mathcal{Z}$, the tail step sizes $\{a_t, b_t : t \ge N_e\}$ still satisfy

$$\sum_{t \ge N_e} a_t = \sum_{t \ge N_e} b_t = \infty, \qquad \sum_{t \ge N_e} (a_t^2 + b_t^2) < \infty, \qquad b_t / a_t \to 0.$$

The martingale-difference and bounded-noise assumptions are inherited by the shifted filtration. Therefore Theorem 4.15, applied with the horizon frozen at L_e and with initial condition $(q_{N_e}^{(e)}, z_{N_e})$, gives

$$\left\|q_t^{(e)} - q_{L_e}^*(z_t)\right\|_\infty \to 0 \qquad \text{a.s.} \tag{69}$$

for the counterfactual process that would continue to run horizon L_e indefinitely. Hence the hitting time

$$\sigma_e(\varepsilon_e) := \inf\left\{t \geq N_e + 1 : \left\|q_t^{(e)} - q_{L_e}^*(z_t)\right\|_\infty \leq \varepsilon_e\right\}$$

is finite almost surely. The state $q_t^{(e)}$ and the slow state z_t are available at the start of generation step t and hence are $\mathcal{F}_{t-1}$-measurable. Therefore $\sigma_e(\varepsilon_e)$ is predictable. Choosing $N_{e+1} = \sigma_e(\varepsilon_e)$ produces an epoch transition satisfying the epoch tracking certificate. Since $N_{e+1} > N_e$, induction gives $N_e < \infty$ for all e on a probability-one event, and $N_e \geq e$ implies $N_e \to \infty$.

This proves only epochwise tracking. It deliberately does not compare $q^{(e)}$ and $q^{(e+1)}$, which live in different spaces when $L_{e+1} > L_e$.

Part (b): fixed-horizon interpretation. During epoch e, all objects entering the fixed-horizon recursion are defined on the finite set $\mathcal{X}_{L_e}$. Thus Theorem 4.15 applies on that epoch, while Lemmas A.4 and A.5 provide the two nonstandard ingredients required for tracking: global stability of the frozen fast ODE and Lipschitz movement of its equilibrium with the slow state. At the transition to epoch $e+1$, the state dimension changes from $\mathbb{R}^{\mathcal{X}_{L_e}}$ to $\mathbb{R}^{\mathcal{X}_{L_{e+1}}}$. The theorem makes no assertion that these finite-dimensional ODEs are restrictions of a single ODE on $\bigcup_e \mathcal{X}_{L_e}$. This proves the stated consistency interpretation.

Part (c): recurrence after the horizon has become large enough. Fix a completed valid program $u \in \mathcal{U}$. Define $e(u) = \min\{e : L_e \geq \ell(u)\}$. Since $L_e \uparrow \infty$, $e(u) < \infty$. By part (a), $\tau_u = N_{e(u)} < \infty$ almost surely. Since the horizons are nondecreasing across epochs, for every $t \geq \tau_u$ the active horizon satisfies $L(t) \geq \ell(u)$. Therefore Lemma 4.3 gives

$$\mathbb{P}(E_t(u) \mid \mathcal{F}_{t-1}) \geq \kappa(u) \qquad \text{a.s. on } \{t \geq \tau_u\}. \tag{70}$$

We now use a stopping-time version of Lemma A.2. For completeness, we spell it out. Since τ_u is a predictable finite epoch-start time, condition on the disjoint events $\{\tau_u = m\}$, each of which is $\mathcal{F}_{m-1}$-measurable. On $\{\tau_u = m\}$, (70) gives

$$\mathbb{P}(E_t(u) \mid \mathcal{F}_{t-1}) \geq \kappa(u) \qquad \text{for all } t \geq m.$$

Applying Lemma A.2 from deterministic time m yields, on $\{\tau_u = m\}$,

$$\mathbb{P}(E_t(u) \text{ i.o.} \mid \mathcal{F}_{m-1}) = 1, \qquad \mathbb{E}[T_u - m \mid \mathcal{F}_{m-1}] \leq \kappa(u)^{-1}.$$

Summing over m gives the almost-sure recurrence claim and the conditional bound (24). Taking expectations also gives $\mathbb{E}[T_u] \leq \mathbb{E}[\tau_u] + \kappa(u)^{-1}$ whenever τ_u is integrable.

Finally, $\mathcal{U}$ is countable because it is a subset of $(\Sigma \cup \{\texttt{EOS}\})^*$ over a finite alphabet. For each fixed u, the recurrence event has probability one; taking the countable intersection over $u \in \mathcal{U}$ gives the simultaneous recurrence claim. This completes the proof. □

B A Concrete Slow-Update Specialization

This section supplies an explicit $h(q, z)$ with verified regularity. It is a finite projected squared-loss learner, not the Bayesian solver used in the main toy and not a general neural-training theorem.

Let Ω_L be finite and $\chi : \Omega_L \to \{1, \ldots, C\}$ a fixed class map. For each outcome u, choose features $x_u \in \mathbb{R}^p$ with $\|x_u\| \leq X$. Let $\Theta = \{\theta : \|\theta\| \leq R\}$, $\mathcal{N} = \Delta_C$, and $z = (\theta, n) \in \mathcal{Z} = \Theta \times \mathcal{N}$. The observation channel is

$$Y \mid U = u, z \sim \text{Bernoulli}(m_z(u)), \qquad m_z(u) = \sigma(v_u^T z + c_u), \quad \|v_u\| \leq B_P,$$

where σ is the logistic function, computed by a numerically stable implementation. Thus $|m_z(u) - m_{z'}(u)| \leq L_P \|z - z'\|$ with $L_P = B_P/4$. A deterministic label $Y(u)$ independent of z is another specialization with $L_P = 0$.

The actual update uses the sampled program and observation:

$$\begin{aligned}\theta_{t+1} &= \Pi_\Theta\{\theta_t - b_t(\theta_t^T x_{U_t} - Y_t)x_{U_t}\},\\ n_{t+1} &= \Pi_{\Delta_C}\{n_t + b_t(e_{\chi(U_t)} - n_t)\}.\end{aligned} \tag{71}$$

This is the stochastic gradient of $\ell(\theta; u, y) = \frac{1}{2}(\theta^T x_u - y)^2$ with a class-frequency state. It is not the gradient of logistic cross entropy even though the observation channel is logistic. Define

$$F(u, y, z) = \begin{bmatrix} -(\theta^T x_u - y)x_u \\ e_{\chi(u)} - n \end{bmatrix}, \qquad h(q, z) = \sum_{u \in \Omega_L} \mu_{q,z}(u) \begin{bmatrix} -(\theta^T x_u - m_z(u))x_u \\ e_{\chi(u)} - n \end{bmatrix}.$$

Assume the reference and mask are fixed and $\alpha_z(s) \geq \alpha_{\min}$ is L_α-Lipschitz. Bounded smooth frontier rewards, smoothed novelty such as clipped $-\log(\epsilon + n_c)$ with $\epsilon > 0$, and bounded Lipschitz information surrogates are examples of rewards satisfying the separate L_r assumption. An arbitrary estimator cache need not satisfy it.

Proposition B.1 (Derived drift, noise, and stability conditions)**.** *On $\mathcal{Q}_L \times \mathcal{Z}$, set*

$$G = (RX + 1)X, \qquad H = \sqrt{G^2 + 4}.$$

The update (71) *stays in the compact set $\mathcal{Z}$. Its unprojected increment is bounded by H, and its centered noise is bounded by $2H$. Moreover,*

$$\begin{aligned}\|h(q, z) - h(q', z')\| \leq \frac{2HL}{\alpha_{\min}}\|q - q'\|_\infty \\ + \left(X^2 + XL_P + 1 + \frac{2HLC_Q L_\alpha}{\alpha_{\min}^2}\right)\|z - z'\|.\end{aligned} \tag{72}$$

Thus this specialization supplies compactness, a Lipschitz drift, and bounded martingale increments required in Assumptions 4.9 and 4.10. The stated step-size and reward regularity assumptions remain necessary.

Proof. Projection enforces $\theta \in \Theta$ and $n \in \Delta_C$ at every step. Since $|\theta^T x - y| \leq RX + 1$ and $\|e_c - n\| \leq 2$, $\|F\| \leq H$. With $U_t \sim \mu_{q_t, z_t}$ and the specified observation law,

$$M_{t+1}^z = F(U_t, Y_t, z_t) - h(q_t, z_t)$$

is $\mathcal{F}_t$-measurable, satisfies $\mathbb{E}[M_{t+1}^z \mid \mathcal{F}_{t-1}] = 0$, and has norm at most $2H$.

To control policy dependence, at each state compare the logits q/α_z and $q'/\alpha_{z'}$:

$$\left\| \frac{q}{\alpha_z(s)} - \frac{q'}{\alpha_{z'}(s)} \right\|_\infty \leq \frac{\|q - q'\|_\infty}{\alpha_{\min}} + \frac{C_Q L_\alpha}{\alpha_{\min}^2}\|z - z'\| =: d.$$

The log-normalizer argument of Lemma A.7 gives path-law TV at most $\tanh(Ld) \leq Ld$. The conditional mean of F at fixed u changes by at most $(X^2 + XL_P + 1)\|z - z'\|$; here the XL_P term accounts for the changing observation kernel and cannot be omitted. Changing the program law contributes at most $2H\mathrm{TV}(\mu_{q,z}, \mu_{q',z'})$, by the dual characterization of the Euclidean norm and boundedness of the vector integrand. Adding these bounds proves (72). □

Clipping bias and numerical approximations. If an implementation clips F to $F^{(c)}$, its decomposition is

$$F^{(c)}(U_t, Y_t, z_t) = h(q_t, z_t) + M_{t+1}^{(c)} + \epsilon_t,$$

where $M_{t+1}^{(c)} = F^{(c)} - \mathbb{E}[F^{(c)} \mid \mathcal{F}_{t-1}]$ is centered and $\epsilon_t = \mathbb{E}[F^{(c)} - F \mid \mathcal{F}_{t-1}]$ is the *expected* clipping bias. A single realized clipping perturbation is not this bias. The SA theorem applies only when the resulting bias vanishes; for example, thresholds at least H make clipping inactive. Other estimation bias must be bounded and included explicitly. This construction derives an admissible finite slow dynamics, not a guarantee that a general neural optimizer, replay summary, or dropout estimator has these properties.

C Experimental Protocol, Additional Results, and Reporting Scope

C.1 Finite environment and observation model

The experiment is a finite Bayesian design problem instantiated with executable Boolean expressions. The hidden hypothesis is one of the 16 truth tables on inputs $(0,0),(0,1),(1,0),(1,1)$. The depth/node-limited DSL enumerates 192 valid program strings at maximum depth 3 and maximum 5 AST nodes, with emitted lengths 2–6 including EOS. Parsing and bounded execution map each string u to a truth table f_u. Multiple strings may have the same semantics; all remain present in the full-grammar experiment. A 64-operation execution budget bounds the interpreter.

For hidden table m, the deterministic specification verifier is $V(u,m) = \mathbf{1}\{f_u = m\}$. The recorded observation is binary, with likelihood

$$p_m(Y=1 \mid u) = \begin{cases} 1-\nu, & f_u = m, \\ \nu, & f_u \neq m, \end{cases} \qquad \nu = 0.02. \tag{73}$$

Thus the noisy channel may disagree with the sound specification check; it is not a sound verifier by itself. The solver maintains log-weights for a finite posterior with uniform initial prior. All hypotheses have common likelihood support. Program evaluation is real bounded DSL execution; the statistical learning target is nevertheless only finite truth-table identification, not unseen-input generalization or neural code synthesis.

Let $b_u = \sum_{m:f_u=m} w(m)$. The predictive success probability is $s_u = \nu + (1-2\nu)b_u$. Binary entropy h_2 gives an exact diagnostic score

$$I_w(u) = h_2(s_u) - h_2(\nu).$$

The implemented online estimator samples $M \sim w$ and $Y \sim p_M(\cdot \mid u)$ and averages $\log[p_M(Y \mid u)/\sum_j w(j)p_j(Y \mid u)]$. It can produce values outside $[0, \log 2]$ before clipping. One semantic-class estimate is copied to every program with that truth table; this is justified here because the observation likelihood and IG coincide within a semantic class. It is not an assumption about arbitrary syntactic aliases in a general environment.

C.2 Control objective and one-round implementation

Write H for the control horizon including any deterministic padding, to distinguish it from an emitted length. The protocol uses depth-scaled row temperatures $\alpha_h = \alpha/\gamma^h$ and a normalized terminal payoff. If the terminal payoff is $F_z(u)$, it can be paid as $F_z(u)/\gamma^{H-1}$ at the last padded transition. Since padding has a singleton action, its KL cost is zero. The trajectory KL chain rule then gives

$$\mathcal{C}_{H,z}(\pi) = \sum_{h<H} \gamma^h \mathbb{E}_\pi[\alpha_h \mathrm{KL}(\pi_h \| \pi_{\mathrm{ref},h})] = \alpha \mathrm{KL}(\mu_\pi \| \rho), \quad \rho = \mu_{\pi_{\mathrm{ref}}}. \tag{74}$$

This equality is specific to the stated temperature/padding normalization; the general theory keeps $\mathcal{C}_{L,z}$ in its statewise form. With normalized non-IG utility o_z and the configured terminal discount Γ, the two evaluation functions are

$$J^{\mathrm{imp}}(\pi) = \mu_\pi^T o_z + \lambda_I \mu_\pi^T(\Gamma \widetilde{I}_z) - \alpha \mathrm{KL}(\mu_\pi \| \rho), \qquad J^{\mathrm{ex}}(\pi) = J^{\mathrm{imp}}(\pi) + \lambda_I \mu_\pi^T[\Gamma(I_z - \widetilde{I}_z)].$$

Exact backward induction finds the surrogate-optimal full policy and the IG-ablated comparator using the same reference and non-IG vector. Neither comparator changes the reference distribution.

The composite non-IG reward favors inexpensive valid programs. The baseline normalized cost averages AST-node count/max-nodes, operation-count/operation-budget, and emitted-length/control-horizon; the utility is one minus that average. A decaying staleness term supplies transient shaping. The pure-IG arm sets the varying cost contribution and this shaping to zero, leaving a constant non-IG payoff. Cost-only and shaping-only arms disable the other term. These are distinct objectives, not estimates of one universal performance ordering.

Frozen round. The model-based primary uses $C_Q = 4$, baseline $\alpha = 0.4$, $\gamma = 0.95$, $\lambda_I = 1$, and relaxation 0.5. Before generation it refreshes the information cache and freezes the posterior, raw/clipped score, non-IG utility, temperature, reference identity, and q_t. It forms π_t from q_t and computes a pending projected Bellman step from the same objective. It samples U_t from that *pre-update* policy, executes it, obtains Y_t, updates the posterior, and installs the pending q_{t+1}. The pending model-based step does not depend on that newly sampled observation. Exact posterior-IG calculations and backward solutions used for audit refer to the frozen w_t, never to w_{t+1}.

This is the deterministic fast-target specialization $M^q_{t+1} = 0$. The posterior may change by $O(1)$ and the finite-run step sizes are not the diminishing two-time-scale schedule of Assumption 4.10. The finite-round certificate does not require that schedule or the conservative invariant-box lower bound on C_Q in the asymptotic theorem. We do not infer asymptotic SA convergence from these runs.

C.3 Estimator budgets, confidence events, and audit quantities

For each of 16 semantic classes, the full-refresh configuration draws 768,700 simulated outcomes per round. The total is 12,299,200 per round, excluding verifier observations used by the learner. For 40 rounds, a trajectory confidence budget 0.05 is split as 0.00125 per round and 7.8125×10^{-5} per refreshed class. With likelihood-log-ratio range bounded by

$$W = 2\log((1-\nu)/\nu),$$

a simultaneous fresh-class radius is $c = W\sqrt{\log(2K_c/\rho_t)/(2n)} \approx 0.02$, where $K_c = 16$. Complete refresh avoids stale-value errors in this configuration. Smaller budgets and imposed estimator distortions are separate tests; they do not inherit the same small radius automatically. Clipping is applied once to the entire raw vector, and the same clipped snapshot is used by selection, updates, and audits.

Exact post-selection information determines whether a stated envelope covered the realized estimator error. A confidence miss is not a deterministic theorem counterexample. The audit separately records deterministic applicability, envelope coverage, signed inequality slack, positivity of the RHS, and numerical tolerance. Failed confidence events must remain in the denominators and in paired outcome statistics. This audit uses known finite-model quantities and is not a prospective certificate available to a general black-box policy.

For each frozen state, define the two RHS diagnostics

$$\begin{aligned} R_t^{\mathrm{exact}} &= \mathcal{D}_{z_t}(\pi^0_{z_t}) + \mathfrak{G}_{z_t}/\lambda_I - \mathcal{E}_{z_t}(\pi^\star_t) - \mathcal{E}_{z_t}(\pi^0_{z_t}) - I_{\max}\varepsilon_t^{\mathrm{tr}}, \\ R_t^{\mathrm{comp}} &= \mathcal{D}_{z_t}(\pi^0_{z_t}) + \mathfrak{G}_{z_t}/\lambda_I - \mathcal{E}_{z_t}(\pi^\star_t) - \mathcal{E}_{z_t}(\pi^0_{z_t}) - I_{\max}B_t^{\mathrm{TV}}. \end{aligned} \tag{75}$$

A positive RHS adds information beyond the universal lower bound zero. By construction $R_t^{\mathrm{comp}} \le R_t^{\mathrm{exact}}$. Reporting both is essential: using exact audited TV does not demonstrate a useful residual-only bound. Exact model errors, declared envelopes, and the Monte Carlo confidence mechanism are distinct quantities; the same symbol is not used for all three.

C.4 Configurations, comparators, and statistical units

The v2 configuration set consists of the composite baseline; pure-IG, cost-only, and shaping-only objectives; a class-mass-balanced reference; three additional temperatures $0.25, 0.15, 0.10$; reference-misalignment magnitudes $0.2, 0.4, 0.6$; projected-step budgets $2, 5, 20$; and injected estimator-bias magnitudes $0.05, 0.10, 0.25$. Every configuration has 40 seeds and 40 observation rounds. Reported selectors include the online projected generator, its oracle-IG projected counterpart, an exact or residual-tracked policy, the IG-ablated comparator, and the raw reference. These selector roles must not be confused with the 17 configuration arms.

The primary local comparisons evaluate candidate policies on the online learner's same frozen state. Whole-rollout comparisons instead run each selector's own posterior process for the matched observation budget. A mean of local IG values across different posteriors is not a test of one-step optimality: a faster learner may have less uncertainty left. It remains legitimate to compare their entropy trajectories and target-identification performance.

Target m is assigned as $s \bmod 16$ for seeds $s = 0, \ldots, 39$: targets 0–7 have three repetitions and targets 8–15 have two. The reported 95% percentile intervals use 5,000 hidden-target-clustered

bootstrap resamples; paired selectors use the same seed identities. They are not confidence intervals based on 1,600 independent rounds. The 17-configuration analysis is exploratory; its intervals do not provide simultaneous familywise coverage or confirm a universal best temperature.

Target success means reaching the experiment's posterior-identification criterion within 40 rounds. The identification criterion is posterior mass at least 0.9 on the true target and a correct MAP decision. The capped time records the first such round, or 41 for a run that does not reach the criterion. This is a bounded event-time diagnostic, not an estimate of an uncensored mean time to success. Entropy AUC is the trapezoidal sum including the initial entropy; semantic coverage counts visited truth tables, and syntactic coverage counts distinct emitted programs. Confidence intervals for the equivalent pure-IG ablated and raw-reference comparisons differ slightly because bootstrap draws differ, so the borderline entropy-AUC significance label is not used as a central claim.

Table 3: Additional same-snapshot information diagnostics (v2). Tilt is $\mathcal{D}(\pi^\star) - \mathcal{D}(\pi^0)$; loss is $\mathcal{D}(\pi^\star) - \mathcal{D}(\pi_t)$. These are means over the reported rounds, not deterministic bounds for each round.

Arm	$\mathcal{D}(\pi_t)$	$\mathcal{D}(\pi^\star)$	$\mathcal{D}(\pi^0)$	Tilt	Loss
Baseline	.03969	.04045	.03040	.01005	.00076
Pure IG	.04159	.04251	.03238	.01012	.00092
Cost only	.03950	.04023	.03008	.01015	.00073
Shaping only	.04128	.04225	.03243	.00982	.00097
Class balanced	.05082	.05280	.04210	.01070	.00198
$\alpha = .25$	.04451	.04585	.02966	.01619	.00134
$\alpha = .15$	.04708	.05012	.02637	.02375	.00304
$\alpha = .10$	.04293	.04927	.02284	.02643	.00634

Table 4: Tracking-budget and estimator-stress audits (v2). Counts are positive information RHS values out of 1,600 online rounds per configuration. An interval containing zero is marked "includes zero," not a demonstrated tie. Bias numbers are configured perturbation magnitudes.

Arm	$R^{\text{exact}} > 0$	$R^{\text{comp}} > 0$	Mean $J^{\text{ex}}(\pi_t) - J^{\text{ex}}(\pi^0)$
Baseline, 1 step	288	0	$-.00003$ (includes zero)
2 steps	447	0	$+.00114$ (positive interval)
5 steps	596	0	$+.00201$ (positive interval)
20 steps	658	0	$+.00217$ (positive interval)
Pure IG	508	0	$+.00257$ (positive interval)
Pure IG, $\alpha = .15$	225	0	$+.00356$ (positive interval)
Bias .05	12	0	$-.00007$ (includes zero)
Bias .10	1	0	$-.00015$ (includes zero)
Bias .25	0	0	$-.00316$ (negative interval)

C.5 Realized-row coverage audit: separate dataset

The v3 coverage-only rerun uses schema 18 and records (17); v2 uses schema 17 and lacks those fields. Table 5 includes the seven configurations with recorded per-arm coverage summaries. No additional configurations are pooled into this table. Each listed configuration has 40 seeds and 40 rounds. The numerical criterion $\kappa_t^{\text{row}}(u) > 10^{-12}$ is a diagnostic threshold, not a theorem of efficient coverage.

The best-covered-program statistic changes little across the reported temperature ladder, while a severely misaligned reference lowers it substantially. This does not prove temperature invariance of the row formula or of every program's probability; both q_t and α_z can change. In particular, an average or retrospective minimum of these floors is not a uniform future floor for adaptive sampling.

Table 5: **Coverage-only rerun (v3):** reported box-floor maximum and realized-row floor summaries. All 192 program floors exceed 10^{-12} in the reported 1,600 rounds for each listed arm, with no audited floor violation. The maximum concerns a best-covered program, not every program. No adaptive future hitting-time claim is inferred from these values.

Arm	Box maximum	Row maximum	Row median
Baseline, $\alpha = .4$	9.85×10^{-19}	.2314	5.47×10^{-4}
Pure IG, $\alpha = .4$	9.85×10^{-19}	.2280	2.21×10^{-3}
$\alpha = .25$	3.72×10^{-29}	.2310	2.08×10^{-3}
$\alpha = .15$	1.10×10^{-47}	.2317	2.08×10^{-3}
$\alpha = .10$	7.55×10^{-71}	.2318	2.29×10^{-3}
Misalignment .6	1.12×10^{-20}	.0040	2.23×10^{-4}
Bias .25	9.85×10^{-19}	.2315	4.58×10^{-4}

C.6 Additional controls and boundaries

The 2-by-2 cost/shaping study reports an entropy-AUC effect relative to the reference of $+5.65$ for the composite baseline, $+5.72$ for cost only, -5.36 for shaping only, and -4.29 for pure IG. Thus disabling cost has a much larger descriptive effect than disabling shaping in that tested design; both terms change in the pure-IG comparison. The class-mass-balanced control retains syntactic aliases and reduces their induced reference-mass imbalance. It is not a canonicalization experiment, and any residual concentration cannot be attributed solely to an intrinsic property of the method without further controls.

An additional expanding-horizon diagnostic reports depth $2 \to 3$, fixed-budget epochs, and preservation of the hypothesis identity and old/new mass partition. Its recorded old-program mass after expansion is about 95.1%, with minimum old-program probability about 1.4×10^{-4}. These figures establish only observed support retention in that run. Its limited target coverage, distinct estimator budget, and incomplete selector metadata prevent pooling it with v2/v3 or using it as evidence for the asymptotic stopping-time theorem.

C.7 Provenance and reporting limits

The numerical tables above report the v2 and v3 experiment summaries. v2 identifies 17 arms, schema 17, and execution-time module hashes; v3 is a separate coverage-only schema. Earlier cross-trajectory objective decompositions were explicitly superseded and are not used here. The reported 9,600-round ordering audit is its designated subset, not a claim that the same denominator covers every arm-selector combination. Neither zero violations nor a positive reported floor empirically proves an infinite-time theorem.

Some descriptive summaries use different emitted versus padded horizon conventions, and the coverage summary count is not uniform across document versions. We therefore do not pool incompatible schemas, quote a conflicting numerical residual multiplier, or infer new aggregate numbers. Reproducing the exact tables requires the per-run configuration manifests, source hashes, trajectory arrays, and bootstrap resample specification; they are not replaced by the prose summaries or by older implementation files. No public release location or independent rerun is asserted here. Full hardware and peak-memory accounting are likewise not inferred from a reported workstation runtime. These limitations constrain replication and scope, not the self-contained finite-round mathematical certificate.

What the tests establish. The finite-model audit checks objective consistency and inequality arithmetic. Shared-snapshot comparisons probe information and KL-cost transfer; separate rollouts test identification under the matched task/channel. Neither tests neural uncertainty calibration or a diminishing-step two-time-scale limit. The measured nonpositive composed RHS values identify a conservative step that still needs improvement. The theory and these observations are reported together without converting a standard identity check into evidence of universally faster self-play.

D SCOPE AND FROZEN-ROUND NOTATION

This part uses the finite syntax-constrained prefix tree, full-support reference, clipped information score, and pre-generation timing of the main text. It proves direct certificates for the same sampling and training algorithm. The finite-round results impose no time-scale separation, no posterior-update smoothness, and no requirement that the projection box contain the optimal value table. The posterior results are a separate, explicitly specified finite-hypothesis specialization. No new empirical results are asserted.

Fix a round and suppress its state index z. The tree has root s_0, at most L actions, and a terminal successor ∂ with value zero. Terminal outcomes retain the complete path, including failures and truncations; their law is never conditioned on acceptance. Write $s^+(s,a)$ for the deterministic successor. Let $0<\gamma<1$ and $\alpha(s)\geq\alpha_{\min}>0$. For a finite table q, define

$$V_q(s)=\alpha(s)\log\sum_a \pi_{\mathrm{ref}}(a\mid s)e^{q(s,a)/\alpha(s)}, \qquad V_q(\partial)=0, \tag{76}$$

$$\pi_q(a\mid s)=\pi_{\mathrm{ref}}(a\mid s)e^{(q(s,a)-V_q(s))/\alpha(s)}, \tag{77}$$

$$d_q(s,a)=q(s,a)-r(s,a)-\gamma V_q(s^+(s,a)), \qquad r_q=\|d_q\|_\infty. \tag{78}$$

The reward r is the actual frozen, once-clipped surrogate reward used by the Bellman operator. For a terminal outcome u, set

$$D_q^{\mathrm{res}}(u)=\sum_{h<\tau(u)}\gamma^h d_q(s_h,a_h), \qquad S_L=\sum_{h=0}^{L-1}\gamma^h. \tag{79}$$

The notation D_q^{res} is a residual return, not the information functional $\mathcal{D}$. For policies π,σ, with σ positive on legal actions, define

$$\mathcal{K}(\pi\|\sigma)=\mathbb{E}_\pi\sum_{h<\tau}\gamma^h\alpha(s_h)\,\mathrm{KL}(\pi(\cdot\mid s_h)\|\sigma(\cdot\mid s_h)). \tag{80}$$

This is a discounted, state-weighted policy divergence. It is not generally a terminal-law KL. Each direction uses the occupancy of its first argument. Write $\mu_q=\mu_{\pi_q}$ and $\mu_0=\mu_{\pi^0}$ for the full terminal laws.

Let $J=J^{\mathrm{imp}}=J_0+\lambda_I\widetilde{\mathcal{D}}$, $J_0=\mathcal{O}-\mathcal{C}$, and $\lambda_I>0$. Let π^0 maximize J_0, and $\pi^\star$ maximize J over all tree policies. Their full-support Gibbs versions are used at all states. All three policies $\pi_q,\pi^0,\pi^\star$ use the same tree, reference, temperature, discount, and reward conventions. Define the implemented non-information gap

$$\Delta_q^0=J_0(\pi^0)-J_0(\pi_q)\geq 0. \tag{81}$$

It differs from the manuscript's ideal-policy gap $\mathfrak{G}_z$. The regularized Bellman and variational identities used here belong to the existing regularized-control framework (Geist et al., 2019); the extension concerns their certificate consequences in this manuscript's setting.

E AN EXACT PERFORMANCE AND INFORMATION IDENTITY

Theorem E.1 (Frozen-policy residual identity)**.** *For every policy π on the finite tree,*

$$J(\pi)=V_q(s_0)-\mathbb{E}_\pi D_q^{\mathrm{res}}-\mathcal{K}(\pi\|\pi_q). \tag{82}$$

Consequently,

$$J(\pi^\star)-J(\pi_q)=\mathcal{K}(\pi_q\|\pi^\star), \tag{83}$$

$$\Delta_q^0=\mathcal{K}(\pi_q\|\pi^0), \tag{84}$$

$$\lambda_I[\widetilde{\mathcal{D}}(\pi_q)-\widetilde{\mathcal{D}}(\pi^0)]=\mathcal{K}(\pi_q\|\pi^0)+\mathcal{K}(\pi^0\|\pi_q)+\mathbb{E}_{\pi^0}D_q^{\mathrm{res}}-\mathbb{E}_{\pi_q}D_q^{\mathrm{res}}. \tag{85}$$

The last identity does not involve $\pi^\star$ or its value table.

Proof. The definition of π_q gives the pointwise equality

$$\alpha(s)\log\frac{\pi_q(a\mid s)}{\pi_{\mathrm{ref}}(a\mid s)} = q(s,a) - V_q(s).$$

Taking an expectation under $\pi(\cdot\mid s)$ yields

$$\mathbb{E}_{a\sim\pi}\left[q(s,a) - \alpha(s)\log\frac{\pi(a\mid s)}{\pi_{\mathrm{ref}}(a\mid s)}\right] = V_q(s) - \alpha(s)\,\mathrm{KL}(\pi_s\|\pi_{q,s}).$$

Insert $q = r + \gamma V_q(s^+) + d_q$. Multiply by γ^h and sum along the episode. The value terms telescope to $V_q(s_0)$ because the terminal value is zero. All sums are finite, including at truncated outcomes. This proves (82); zeros of π are interpreted by $0\log 0 = 0$. Apply the identity with the optimal full-objective value $q^\star$, whose residual is zero, to obtain (83). Apply the same argument to the non-information reward and its optimal table to obtain (84). Finally subtract (82) for π_q and π^0:

$$J(\pi_q) - J(\pi^0) = \mathbb{E}_{\pi^0} D_q^{\mathrm{res}} - \mathbb{E}_{\pi_q} D_q^{\mathrm{res}} + \mathcal{K}(\pi^0\|\pi_q).$$

The left side also equals $-\Delta_q^0 + \lambda_I[\widetilde{\mathcal{D}}(\pi_q) - \widetilde{\mathcal{D}}(\pi^0)]$. Use (84) and rearrange. □

Corollary E.2 (Ideal information tilt as a two-direction divergence)**.** *For the surrogate-optimal policy,*

$$\lambda_I[\widetilde{\mathcal{D}}(\pi^\star) - \widetilde{\mathcal{D}}(\pi^0)] = \mathcal{K}(\pi^\star\|\pi^0) + \mathcal{K}(\pi^0\|\pi^\star). \tag{86}$$

The right side is strictly positive if the two policies differ at a legal reachable state. This statement concerns the clipped surrogate; the exact information comparison requires the error correction below.

Proof. Set $q = q^\star$ in (85). The residual vanishes. Both policies have positive probability at every legal reachable state and every legal action. A difference at one such state makes at least one positively weighted row KL positive. □

F Linear, signed, and quadratic residual certificates

Proposition F.1 (Signed and uniform linear bounds)**.** *Let* $m_q = \min_{u\in\Omega_L} D_q^{\mathrm{res}}(u)$ *and define*

$$\Xi_q^{\mathrm{sign}} = \mathbb{E}_{\pi_q} D_q^{\mathrm{res}} - m_q.$$

Then

$$0 \le J(\pi^\star) - J(\pi_q) \le \Xi_q^{\mathrm{sign}} \le 2S_L r_q. \tag{87}$$

Moreover, writing $\omega_q = \max_u D_q^{\mathrm{res}}(u) - \min_u D_q^{\mathrm{res}}(u)$ *and* $v_{q0} = \mathrm{TV}(\mu_q,\mu_0)$*,*

$$\left|\mathbb{E}_{\pi_q} D_q^{\mathrm{res}} - \mathbb{E}_{\pi^0} D_q^{\mathrm{res}}\right| \le \omega_q v_{q0} \le 2S_L r_q v_{q0}. \tag{88}$$

Proof. Subtract the residual identity for $\pi^\star$ and π_q and discard the nonnegative $\mathcal{K}(\pi^\star\|\pi_q)$. The result is at most $\mathbb{E}_{\pi_q} D_q^{\mathrm{res}} - m_q$. Since $|D_q^{\mathrm{res}}(u)| \le S_L r_q$, the first bound follows. For the second, subtract the minimum of D_q^{res} before applying $|\mathbb{E}_P f - \mathbb{E}_Q f| \le (\max f - \min f)\,\mathrm{TV}(P,Q)$. □

The minimum, maximum, and expectation of the residual return can be obtained by a backward min/max/expectation recursion on the deterministic tree. This still requires the finite model; it is not a model-free guarantee.

For sharper local bounds let $e(s,a) = q^\star(s,a) - q(s,a)$. Choose residual intervals $d^-(s,a) \le d_q(s,a) \le d^+(s,a)$. Starting at the leaves, define

$$\ell(s,a) = -d^+(s,a) + \gamma \min_b \ell(s^+,b), \tag{89}$$

$$u(s,a) = -d^-(s,a) + \gamma \max_b u(s^+,b), \tag{90}$$

where both successor contributions equal zero at ∂. Set

$$b_q(s) = \max_a u(s,a) - \min_a \ell(s,a), \qquad \Xi_q^{\mathrm{quad}} = \mathbb{E}_{\pi_q}\sum_{h<\tau}\gamma^h \frac{b_q(s_h)^2}{8\alpha(s_h)}. \tag{91}$$

At a singleton-action row one may set $b_q(s) = 0$, since its policy KL is zero.

Theorem F.2 (Quadratic, policy-relevant residual bound)**.** *On any event where the residual intervals hold simultaneously,*

$$0 \le J(\pi^\star) - J(\pi_q) \le \Xi_q^{\mathrm{quad}}. \tag{92}$$

In particular, exact residuals satisfy

$$J(\pi^\star) - J(\pi_q) \le \frac{r_q^2}{2\alpha_{\min}} H_L, \qquad H_L = \sum_{h=0}^{L-1} \gamma^h S_{L-h}^2. \tag{93}$$

Thus the information certificate can use

$$\Xi_q = \min\left\{ \Xi_q^{\mathrm{sign}}, \Xi_q^{\mathrm{quad}}, 2S_L r_q, \frac{H_L r_q^2}{2\alpha_{\min}} \right\}. \tag{94}$$

Only the candidates actually justified by available exact quantities or valid confidence bounds are included in this minimum.

Proof. Weighted log-sum-exp obeys

$$\min_b e(s,b) \le V_{q^\star}(s) - V_q(s) \le \max_b e(s,b).$$

The fixed-point equation gives $e(s,a) = -d_q(s,a) + \gamma[V_{q^\star}(s^+) - V_q(s^+)]$. Backward induction therefore gives $\ell \le e \le u$ and $\mathrm{span}_a\, e(s,a) \le b_q(s)$.

For completeness, if a bounded random variable X has range width b, let $F(t) = \log \mathbb{E}e^{tX}$. Under exponential tilting, $F''(t) = \mathrm{Var}_t(X) \le b^2/4$: centering at the interval midpoint bounds the second moment by $b^2/4$, and variance is no larger. Integrating twice gives $F(1) - F(0) - F'(0) \le b^2/8$. At row s take $X = e(s,A)/\alpha(s)$ under $A \sim \pi_q(\cdot \mid s)$. Since $\pi^\star \propto \pi_q e^{e/\alpha}$,

$$\mathrm{KL}(\pi_{q,s} \| \pi_s^\star) = \log \mathbb{E}_{\pi_{q,s}} e^{e/\alpha} - \mathbb{E}_{\pi_{q,s}}(e/\alpha) \le \frac{b_q(s)^2}{8\alpha(s)^2}.$$

Multiply by discounted occupancy and $\alpha(s)$ and use (83). At depth h, backward induction also yields $|e(s,a)| \le S_{L-h} r_q$. Hence its row span is at most $2S_{L-h} r_q$. Summing the resulting bounds gives (93). The minimum of valid upper bounds remains valid. □

Remark F.3 (Invariance to irrelevant value offsets). Replace $q(s,a)$ by $q(s,a) + c(s)$, with $c(\partial) = 0$. The policy is unchanged and $d_q(s,a)$ changes by $c(s) - \gamma c(s^+)$. Its path return changes only by $c(s_0)$. Hence Ξ^{sign}, ω_q, and the residual expectation difference are unchanged. For exact intervals in (90), both endpoints shift by $-c(s)$, so $b_q(s)$ and Ξ^{quad} are also unchanged. This invariance need not hold for the coarse sup-norm residual bounds.

G CONTRAST-AWARE INFORMATION GUARANTEES

The manuscript's information envelope implies $|\widetilde{I}(x) - I(x)| \le \delta(x)$ after clipping. Let $\Gamma(u) \in [0,1]$ be the exact discount/normalization entering the information functional, and define

$$E_{q0}^{\Delta} = \sum_{u \in \Omega_L} |\mu_q(u) - \mu_0(u)| \Gamma(u) \delta(x(u)). \tag{95}$$

This is no larger than $\mathcal{E}(\pi_q) + \mathcal{E}(\pi^0)$. For a uniform envelope $\delta \le \delta_{\max}$ it is at most $2\delta_{\max}\, \mathrm{TV}(\mu_q, \mu_0)$. If errors are known to be identical within a semantic class, sum the signed coefficients $(\mu_q - \mu_0)\Gamma$ within that class before taking absolute values. That further refinement requires genuinely shared errors, not merely a proposed semantic similarity.

Theorem G.1 (Implemented-comparator information certificate)**.** *Under the frozen-round assumptions and a valid simultaneous information error envelope, let Ξ_q be any upper bound on $J(\pi^\star) - J(\pi_q)$. Define*

$$M_q^{\mathrm{obj}} = \frac{\Delta_q^0 - \Xi_q}{\lambda_I} - E_{q0}^{\Delta}, \tag{96}$$

$$M_q^{\mathrm{sym}} = \frac{\Delta_q^0 + \mathcal{K}(\pi^0 \| \pi_q) - \omega_q v_{q0}}{\lambda_I} - E_{q0}^{\Delta}, \tag{97}$$

$$M_q = \max\{M_q^{\mathrm{obj}}, M_q^{\mathrm{sym}}\}. \tag{98}$$

Then

$$\mathcal{D}(\pi_q) - \mathcal{D}(\pi^0) \geq M_q. \tag{99}$$

In particular $M_q > 0$ certifies an information advantage over the IG-ablated comparator on the same state, not merely a positive absolute information value. The completed-path support floor of the manuscript continues to hold for the same q, without modification.

Proof. Optimality gives $J(\pi_q) \geq J(\pi^\star) - \Xi_q \geq J(\pi^0) - \Xi_q$. Expand $J = J_0 + \lambda_I \widetilde{\mathcal{D}}$ to obtain $\lambda_I(\widetilde{\mathcal{D}}_q - \widetilde{\mathcal{D}}_0) \geq \Delta_q^0 - \Xi_q$. Alternatively use (85) and (88) to obtain its symmetric counterpart. For either inequality the conversion to exact information costs at most

$$\left| \sum_u [\mu_q(u) - \mu_0(u)]\Gamma(u)[\widetilde{I}(x(u)) - I(x(u))] \right| \leq E_{q0}^{\Delta}.$$

Both bounds hold on the same event, hence so does their maximum. The path-floor argument uses only bounded table entries, positive reference probabilities, and a positive temperature floor; none has changed. □

Absolute information and what can be computed. Set $I^-(x) = \max\{0, \widetilde{I}(x) - \delta(x)\}$ and $\underline{D}_0 = \sum_u \mu_0(u)\Gamma(u)I^-(x(u))$. Then a bounded, nonnegative absolute certificate is

$$L_q = \max\{0, \min\{I_{\max}, \underline{D}_0 + M_q\}\} \leq \mathcal{D}(\pi_q). \tag{100}$$

No ideal full-objective policy appears in (98), but model-based expectations, the exact ablated comparator, and residual bounds are still required. If the *discounted functional* $\mathcal{O}(\pi)$ is constant over all admissible policies (for example, if non-IG rewards vanish), then $\pi^0 = \pi_{\mathrm{ref}}$. A constant terminal reward alone is insufficient when termination lengths differ. With a general non-IG utility, computing π^0 remains a control problem. This is not an oracle-free black-box certificate in the usual computational sense; it only removes the full-objective optimal-policy oracle from the displayed bound.

Best direct interval audit on a fully enumerable model. Let $I^+(x) = \min\{I_{\max}, \widetilde{I}(x) + \delta(x)\}$ and $a_u = [\mu_q(u) - \mu_0(u)]\Gamma(u)$. Directly,

$$\mathcal{D}(\pi_q) - \mathcal{D}(\pi^0) \geq \sum_{a_u \geq 0} a_u I^-(x(u)) + \sum_{a_u < 0} a_u I^+(x(u)). \tag{101}$$

For unrestricted independent intervals this is the sharp lower bound over the interval box: minimize each linear term at the appropriate endpoint. Shared semantic errors call for aggregation first. If the complete terminal law and all intervals are available, this direct audit must be reported too; the residual theorem should not be presented as computationally necessary. Indeed, denote the right side of (101) by B_q^{int}. On the same envelope event,

$$B_q^{\mathrm{int}} \geq \widetilde{\mathcal{D}}(\pi_q) - \widetilde{\mathcal{D}}(\pi^0) - E_{q0}^{\Delta} \geq M_q. \tag{102}$$

For a nonnegative coefficient, $I^- \geq \widetilde{I} - \delta$; for a negative coefficient, $I^+ \leq \widetilde{I} + \delta$. Multiplication by the coefficient in each case and summation prove the first inequality. The residual bounds proved above give the second. Thus the residual certificate is not tighter than direct interval comparison when all laws and intervals are available. Its role is to expose the optimization/error trade-off and give guarantees from certified residual summaries; no additional computational advantage is asserted without a specified information-access model.

H Non-vacuity conditions and a fully specified construction

Corollary H.1 (A checkable residual threshold)**.** *Let $\theta_q = \Delta_q^0 - \lambda_I E_{q0}^{\Delta} > 0$. A sufficient condition for strict same-state information advantage is either*

$$r_q < \frac{\theta_q}{2S_L} \quad \textit{or} \quad r_q < \sqrt{\frac{2\alpha_{\min}\theta_q}{H_L}}. \tag{103}$$

The local quadratic or signed bound can replace its uniform counterpart and may certify a larger region. These are statewise, a posteriori criteria: Δ_q^0 and E_{q0}^{Δ} also depend on q.

Proof. Each condition makes the corresponding upper bound Ξ_q smaller than θ_q, whence $M_q^{\mathrm{obj}} > 0$. □

Proposition H.2 (A nonempty neighborhood with a positive certificate)**.** *Fix the finite model, temperature, and surrogate. Suppose $\pi^\star \neq \pi^0$ and let $\Delta_\star^0 = \mathcal{K}(\pi^\star \| \pi^0) > 0$. If $\lambda_I E_{\star 0}^{\Delta} < \Delta_\star^0$, using exact residuals in the uniform quadratic certificate gives $M_q^{\mathrm{obj}} > 0$ throughout some neighborhood of $q^\star$. In particular this holds for exact information scores and $\pi^\star \neq \pi^0$.*

Proof. On a fixed finite tree, softmax policies, terminal laws, Δ_q^0, and E_{q0}^{Δ} are continuous in q. Also $r_q \le (1+\gamma)\|q - q^\star\|_\infty$, so the quadratic penalty tends to zero. At $q = q^\star$ the strict margin assumption holds. Continuity preserves it on an open neighborhood. This proves existence, not that a tested trajectory enters that neighborhood or that its projection box allows it to do so. □

Two-query construction. Take two equiprobable hypotheses $M \in \{0, 1\}$ and two completed programs, each consisting of a root token followed by EOS, so $L = 2$. For query 0, Y is an independent fair bit; for query 1, $Y = M$ flipped with probability $\nu = 0.02$. The exact information scores are

$$I_0 = 0, \qquad I_1 = \log 2 - h_2(0.02) =: I_\bullet.$$

Use terminal information reward only, $\gamma = 0.95$, $\lambda_I = 1$, $\alpha(s) = 0.4$, and reference probability $1/2$ for each root token. All EOS rows are singleton rows. The optimum has $q^\star(s_0, 0) = 0$, $q^\star(s_0, 1) = \gamma I_\bullet$, and EOS values $0, I_\bullet$. Set the implemented root values to $(0.03, \gamma I_\bullet - 0.03)$ and leave EOS values exact. All values lie in $[-4, 4]$ and $r_q = 0.03$. At the root $b_q = 0.06$; singleton rows contribute zero. Therefore

$$\Xi_q^{\mathrm{quad}} = 0.06^2/(8 \cdot 0.4) = 0.001125.$$

Even with the declared common information envelope $\delta = 0.02$, the direct advantage certificate is positive (the numerical values are reproduced by the supplied `verification/independent_checks.py`; this script implements synthetic arithmetic checks, not the v2/v3 experiments). The manuscript's coarse TV substitution is $\tanh(2 \cdot 0.03/(0.4 \cdot 0.05)) = \tanh 3$ and its absolute RHS is negative. This is a specified mathematical example in the model class, not a rerun of the manuscript's 192-program experiment.

Quantity in this construction	Value
True exact-information advantage	0.158077
New objective-residual advantage lower bound	0.054547
New symmetric-contrast advantage lower bound	0.113899
Original residual-TV absolute-information lower bound	−0.365538

Terminal-KL specialization. Only when the objective actually reduces to $\mu^T o + \lambda_I \mu^T f - \alpha\,\mathrm{KL}(\mu\|\rho)$, with $f = \Gamma\widetilde{I}$, let $\mu_0 \propto \rho e^{o/\alpha}$ and $\mu_s \propto \mu_0 e^{sf/\alpha}$. Direct differentiation gives

$$\frac{d}{ds}\mathbb{E}_{\mu_s} f = \frac{1}{\alpha}\,\mathrm{Var}_{\mu_s}(f), \qquad \widetilde{\mathcal{D}}(\pi^\star) - \widetilde{\mathcal{D}}(\pi^0) = \frac{1}{\alpha}\int_0^{\lambda_I} \mathrm{Var}_{\mu_s}(f)\,ds.$$

Indeed differentiation of each normalized exponential weight gives $\frac{d}{ds}\mu_s(u) = \mu_s(u)(f(u) - \mathbb{E}_{\mu_s} f)/\alpha$. Thus the ideal surrogate advantage is strictly positive when f is nonconstant. This specialization is not silently applied to the general state-dependent discounted objective.

I Expected Bayesian entropy budget and confidence scope

Corollary I.1 (Direct-certificate entropy budget)**.** *For the manuscript's coherent finite Bayesian learner, suppose the information and residual envelopes used above hold almost surely and let L_t denote* (100) *at the frozen pre-generation state. Then*

$$\mathbb{E}[H(w_t) - H(w_{t+1}) \mid \mathcal{F}_{t-1}] \ge L_t, \qquad \mathbb{E}H(w_{T+1}) \le H(w_1) - \sum_{t=1}^{T} \mathbb{E}L_t. \tag{104}$$

This has no dependence on an ideal full-objective policy.

Proof. The mixture-predictive identity gives expected entropy reduction $d_t = \mathbb{E}_{\mu_t} I_{z_t}(x(U_t))$. Since $I \geq 0$ and $0 \leq \Gamma \leq 1$, $d_t \geq \mathcal{D}_{z_t}(\pi_t) \geq L_t$. Taking expectations and telescoping proves the claim. □

Validation and failures. If $|d_q - \widehat{d}| \leq c(s,a)$ holds simultaneously after the policy snapshot is fixed, use $d^{\pm} = \widehat{d} \pm c$ in (90). A uniform radius c gives $r_q \leq \widehat{r} + c$ in the coarse bounds. The exact signed bound and residual-contrast term cannot use an estimated residual without its own error accounting. One safe choice is to omit these candidates and retain the uniform or interval quadratic upper bound. Another is to propagate interval min/max bounds and bound both residual expectations. No validation sample may silently be reused to select an arbitrary new policy under a merely fixed-policy guarantee.

For repeated audits, budgets ρ_t with $\sum_t \rho_t \leq \rho$ give a simultaneous event by the union bound. This event does not justify conditioning the Bayesian dynamics on all future successful audits. For an unconditional entropy budget, let $\mathcal{H}_t$ be the history just before the fresh estimator/validation randomness and let A_t be its joint coverage event with $\mathbb{P}(A_t^c \mid \mathcal{H}_t) \leq \rho_t$. The pregeneration sigma-field includes those fresh draws. Because $0 \leq L_t \leq I_{\max}$ and $d_t \geq 0$ always, $d_t \geq L_t - I_{\max}\mathbf{1}_{A_t^c}$. Thus

$$\mathbb{E}H(w_{T+1}) \leq H(w_1) - \sum_{t=1}^{T} \mathbb{E}L_t + I_{\max} \sum_{t=1}^{T} \rho_t. \tag{105}$$

Internal simulations must not reveal the latent hypothesis beyond the observed history. The predictable confidence guarantee must include all estimation channels used in L_t.

J Coverage produces finite-hypothesis learning

This section is under a fixed true hypothesis, not under a mixture-predictive sampling law. It therefore complements, rather than reinterprets, the Bayesian entropy identity. Standard controlled-sensing theory studies adaptive query selection (Nitinawarat et al., 2013); no claim of a new general hypothesis-testing principle is made here.

Assume $K \geq 2$ hypotheses, positive prior $w_1(m) > 0$, and specified finite observation kernels $p_m(y \mid u)$ with common positive support for each query. Fix the true index i. Under $\mathbb{P}_i$, the observation at round t, conditional on the history and $U_t = u$, follows $p_i(\cdot \mid u)$. The generator and its internal randomness use only observed history. Posterior updates use these same likelihoods. No bounded log-likelihood ratio, time-scale separation, or independence of the chosen queries is needed. Define affinity and distinguishing strength

$$a_{ij}(u) = \sum_y \sqrt{p_i(y \mid u) p_j(y \mid u)}, \qquad b_{ij}(u) = 1 - a_{ij}(u) \in [0,1]. \tag{106}$$

Use any predictable pointwise floor $0 \leq k_t(u) \leq \mu_t(u)$, for example the realized-row floor for completed paths and zero for other outcomes. Set

$$\underline{h}_{t,ij} = \sum_u k_t(u) b_{ij}(u), \qquad A_{T,ij} = \sum_{t=1}^{T} \underline{h}_{t,ij}. \tag{107}$$

One may sum over a finite selected set of witness programs only; omitted outcomes contribute zero. For an unknown true hypothesis, a conservative observable choice is $\underline{h}_t = \min_{i \neq j} \underline{h}_{t,ij}$, computed from the specified finite model.

Theorem J.1 (Anytime posterior certificate from adaptive row floors)**.** *For every wrong hypothesis $j \neq i$, the process*

$$Z_{T,ij} = \sqrt{\frac{w_{T+1}(j)/w_{T+1}(i)}{w_1(j)/w_1(i)}}\, e^{A_{T,ij}}, \qquad Z_{0,ij} = 1, \tag{108}$$

is a nonnegative supermartingale under $\mathbb{P}_i$. For any $0 < \delta < 1$, with probability at least $1 - \delta$, simultaneously for all $T \geq 0$ and $j \neq i$,

$$\frac{w_{T+1}(j)}{w_{T+1}(i)} \leq \frac{w_1(j)}{w_1(i)} \left(\frac{K-1}{\delta}\right)^2 e^{-2A_{T,ij}}. \tag{109}$$

If $A_{T,ij} \to \infty$ almost surely for every $j \neq i$, then $w_t(i) \to 1$ and $H(w_t) \to 0$ almost surely.

Proof. Let $O_{t,j} = w_{t+1}(j)/w_{t+1}(i)$ and $O_{0,j} = w_1(j)/w_1(i)$. Bayes' rule gives $O_{t,j}/O_{t-1,j} = p_j(Y_t \mid U_t)/p_i(Y_t \mid U_t)$. Conditional on the history, including the frozen internal randomization,

$$\mathbb{E}_i\left[\sqrt{O_{t,j}/O_{t-1,j}} \mid \mathcal{F}_{t-1}\right] = \sum_u \mu_t(u)a_{ij}(u) = 1 - \sum_u \mu_t(u)b_{ij}(u) \leq 1 - \underline{h}_{t,ij}.$$

Because $e^x(1-x) \leq 1$ for $0 \leq x \leq 1$, multiplication by the predictable factor $e^{\underline{h}_{t,ij}}$ proves the supermartingale property. Positivity of the kernels keeps all odds well defined.

For a nonnegative supermartingale starting at one, stopping at the first crossing of $c > 0$ and at a deterministic n gives $c\mathbb{P}(\max_{T \leq n} Z_T \geq c) \leq 1$. Let $n \to \infty$ and choose $c = (K-1)/\delta$. A union bound over $j \neq i$, followed by squaring and rearranging, yields (109). For any fixed $\delta > 0$, on the event in (109), divergence of every $A_{T,ij}$ forces $O_{T,j} \to 0$. Since the probability of this event is at least $1 - \delta$ for every $\delta > 0$, the convergence holds almost surely. There are finitely many wrong hypotheses, so $w_{T+1}(i) = 1/(1 + \sum_{j \neq i} O_{T,j}) \to 1$. Entropy is continuous on the finite simplex, proving its convergence to zero. □

Corollary J.2 (Deterministic finite-time and mean hitting-time bounds)**.** *Suppose $\underline{h}_{t,ij} \geq h_{t,ij} \geq 0$ with deterministic lower bounds. Define $B_i = \sum_{j \neq i} \sqrt{w_1(j)/w_1(i)}$. For $0 < \epsilon < 1$,*

$$\mathbb{P}_i\{w_{T+1}(i) < 1 - \epsilon\} \leq \sqrt{\frac{1-\epsilon}{\epsilon}} \sum_{j \neq i} \sqrt{\frac{w_1(j)}{w_1(i)}} e^{-\sum_{t=1}^T h_{t,ij}}. \tag{110}$$

In particular, if all $h_{t,ij} \geq h > 0$, the right side is $B_i\sqrt{(1-\epsilon)/\epsilon}\, e^{-hT}$. A sufficient fixed-round budget for failure probability at most δ is

$$T \geq \frac{1}{h}\left[\log\frac{B_i}{\delta} + \frac{1}{2}\log\frac{1-\epsilon}{\epsilon}\right]_+. \tag{111}$$

Round up to an integer. If τ_ϵ is the first observed round at which the posterior mass on i is at least $1 - \epsilon$, then, under an indefinitely continued process,

$$\mathbb{E}_i\tau_\epsilon \leq 1 + \frac{\log^+(B_i\sqrt{(1-\epsilon)/\epsilon}) + 1}{h}. \tag{112}$$

For $\epsilon < 1/2$, the posterior-mass event also ensures a correct unique MAP.

Proof. The one-step calculation in the previous proof and iteration give $\mathbb{E}_i\sqrt{O_{T,j}} \leq \sqrt{O_{0,j}}\prod_t(1 - h_{t,ij}) \leq \sqrt{O_{0,j}}e^{-\sum_t h_{t,ij}}$. If $w_{T+1}(i) < 1-\epsilon$, then $\sum_j O_{T,j} > \epsilon/(1-\epsilon)$, and hence $\sum_j \sqrt{O_{T,j}} > \sqrt{\epsilon/(1-\epsilon)}$. Markov's inequality proves (110). The sample budget follows by rearrangement. For hitting time and $T \geq 1$, $\{\tau_\epsilon > T\} \subseteq \{w_{T+1}(i) < 1 - \epsilon\}$. Use $\mathbb{E}\tau_\epsilon = 1 + \sum_{T \geq 1}\mathbb{P}(\tau_\epsilon > T)$ and bound $\sum_{T \geq 1}\min\{1, Ae^{-hT}\}$ by the integral on $[0, \infty)$, which is at most $(\log^+ A + 1)/h$. The final assertion follows from $w(i) > 1/2$. □

How the manuscript's coverage assumptions supply h. For each pair $i \neq j$, suppose an eligible completed witness u_{ij} has $b_{ij}(u_{ij}) > 0$. The time-uniform box floor yields $\underline{h}_{t,ij} \geq \kappa(u_{ij})b_{ij}(u_{ij}) > 0$. A finite collection of witnesses therefore gives a positive common h and the preceding bounds. With expanding horizons, start after the witnesses become eligible: conditional bounds use the posterior at that predictable start. Almost-sure consistency survives an almost-sure finite start, but an unconditional expected runtime additionally needs control of that start. The constants can still be exponentially poor in program length. Realized-row floors improve the *accumulated* anytime certificate in (109); the inverse of one current floor is still not an adaptive future waiting-time guarantee.

Specialization to the noisy equality-query toy. For different truth tables i, j, a query whose semantics equal i or j has $b_{ij}(u) = 1 - 2\sqrt{\nu(1-\nu)}$; all other queries have $b_{ij}(u) = 0$. At $\nu = 0.02$ this distinguishing strength is exactly 0.72. Writing $k_t(c)$ for the sum of program floors in semantic class c,

$$\underline{h}_{t,ij} = 0.72[k_t(i) + k_t(j)]. \tag{113}$$

If $k_t(c) \geq k_{\min} > 0$ for every class and round, one may take $h = 1.44k_{\min}$. A uniform 16-hypothesis prior gives $B_i = 15$. This is a fixed-target theorem and does not require drawing the benchmark target from the prior. It does require the specified likelihood to be correct.

K What these results do not imply

Proposition K.1 (No uniform positive information or universal improvement)**.** *The manuscript's basic support and boundedness assumptions alone imply neither a strictly positive information advantage over π^0 nor posterior identification. No coherent finite Bayesian learner can have a strictly positive constant expected entropy decrease at every round indefinitely.*

Proof. Take two hypotheses with identical observation kernels at every query. Then $I \equiv 0$ and every affinity is one. All support assumptions can hold, but posterior odds never change and information advantages are zero. For the last claim, telescoping entropy gives $\sum_{t=1}^{T} \mathbb{E}[H(w_t) - H(w_{t+1})] = H(w_1) - \mathbb{E}H(w_{T+1}) \leq H(w_1) \leq \log K$. A positive constant per-round lower bound contradicts this as $T \to \infty$. □

Same-state information advantage is not a theorem comparing two complete adaptive rollouts: their future posterior states differ. The posterior coverage theorem establishes learnability under identifiability, but does not establish that IG is necessary or faster than persistent reference sampling. General neural calibration, neural optimization, nonmyopic curriculum optimality, and minimax-optimal sample complexity are not proved. The sharper certificates do not determine how many positive rounds occur in v2 or v3 without the frozen trajectory records.

AI use statement

Generative AI tools, including ChatGPT Codex Claude(code), assisted in critiquing and polishing the theoretical framework, checking proofs and refining hypotheses, designing and reviewing experiments, writing and checking implementation and verification code, and interpreting experimental reports. AI assistance also covered literature discovery and summarization. AI-assisted code was used in developing the finite synthetic experiment and the numerical verification routines; the stated experimental measurements come from program execution, not from generated prose.

The checking process included tracing claims to the supplied experimental records, reviewing assumptions and proof dependencies, independently implementing finite-tree arithmetic checks, and checking citations, cross-references, compilation, and rendered pages. These checks are not machine-verified proofs and do not constitute an independent rerun of the v2/v3 experiments. The direct certificates introduced here have not been evaluated on those trajectories. The authors retain responsibility for the final text, mathematical claims, code, and reported results, including all AI-assisted material.

Ethics statement

The study concerns a bounded Boolean DSL and synthetic finite hypotheses, not human participants or deployment of an unrestricted code-generating agent. For larger implementations, independent sandboxing, resource limits, and audits of generated tasks would be necessary. A verifier certifies only its stated specification; syntax support and noisy observations are not safety guarantees.

Reproducibility statement

The model, decoder, and algorithm are specified in Sections 3 and 4. Complete proofs and the explicit slow-update specialization appear in Sections A, B, D to G and J. Section C specifies the

finite environment, observation model, estimator budget, comparisons, and statistical units. The accompanying verification script reproduces the two-query construction and checks the new identities on synthetic finite trees. It does not reproduce the v2/v3 experiments: their trajectory archives and resampling manifests are not included in this source package. This distinction and remaining empirical reproducibility limits are recorded in Section C.7.