UNDERSTANDING SINGLE- AND MULTI-AGENT LEARNING: COMPLEXITY, ALIGNMENT, AND ERROR PROPAGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

When does decomposing a task across specialized agents improve learning efficiency? Empirical studies report mixed outcomes across tasks, training procedures, and resource budgets. We develop a comparative learning-theoretic framework for monolithic and modular sequential policies, organizing established tools around three mechanisms: statistical class complexity, objective and policy-class mismatch, and stagewise learning errors. Under conditional-value access, independent policy and reward factorization yields a local-complexity guarantee; dependent modular policies can match the single-agent sufficient budget under matched dimension, horizon, and radius allocations. Reward discrepancy and class-optimum gaps qualify these comparisons, while a separate reference-demonstration analysis relates local learning rates to downstream error propagation. Finite-class and exact Gaussian regression examples show why neither agent count nor dependency graph alone determines learning costs. The framework makes explicit the observation models and comparators needed to attribute efficiency differences to decomposition. Its PAC results compare sufficient budgets, while the regression example gives exact expected excess risks under stated assumptions.

1 INTRODUCTION

Should a single agent learn an entire task, or should specialized agents learn its constituent parts? This choice is central to LLM-based systems that combine planning, reasoning, and verification (Tran et al., 2025; Guo et al., 2024). Decomposition can simplify individual learning problems, but it can also change the feasible policies, training objectives, and consequences of upstream errors. Understanding these effects is necessary for evaluating the learning efficiency of single-agent (SARL) and modular multi-agent (MARL) formulations.

A practical question with mixed empirical evidence. The broader literature reports different outcomes for single-agent and multi-agent systems. Optima improves task performance and token efficiency through multi-agent training (Chen et al., 2025). Controlled evaluations find coordination gains that vary with task structure (Kim et al., 2026), while budget-aware multi-hop reasoning evaluations find single-agent systems competitive with, and often better than, multi-agent alternatives (Tran & Kiela, 2026). These studies use different tasks, training procedures, and resource measures. Their findings motivate a conditional understanding of decomposition rather than a universal architectural ranking. We focus on the learning side of this question: **under what conditions does task decomposition reduce the samples sufficient to learn a good policy?** Training-sample requirements and inference-token budgets are distinct quantities.

Why a comparative framework is needed. Relevant tools exist in statistical learning, structured RL, and sequential prediction, but their guarantees answer different questions. A class-relative generalization bound does not account for changing the training objective or attainable policies. An execution-error bound requires a learning rate before it yields a sample budget. Comparisons also depend on whether a sample supplies conditional values, a complete reference demonstration, or interactive feedback. Organizing these results therefore requires specifying the observation model, objective, policy class, and performance comparator before comparing rates. Otherwise, a benefit

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.

from a restricted class or an error from a particular training protocol can be attributed to agent count alone.

Contribution. We develop a **comparative learning-theoretic framework** for monolithic and modular sequential policies. The contribution is the systematic separation of three mechanisms and the conditions this reveals for beneficial decomposition. We organize established PAC, reward-transfer, and sequential-prediction arguments around the practical decomposition question (Valiant, 1984; Ross & Bagnell, 2010; Ross et al., 2011), specifying what each comparison holds fixed and how its conclusions depend on the observation model:

• **Statistical complexity of decomposition.** We derive and compare sufficient prompt budgets under conditional-value access. Independent policies, local rewards, and Cartesian trainable parameters yield a guarantee governed by the largest local complexity, with a logarithmic confidence cost for the number of agents. Fully history-dependent policies can match the SARL certificate under matched total dimension, equally divided horizons, and equal parameter radii (Theorems 3.1–3.3; Propositions 3.4–3.5). These comparisons identify when specialization reduces the statistical budget and show why sequential dependence alone need not enlarge the joint certificate.

• **Objective and policy-class mismatch.** We account separately for reward distortion and differences in attainable performance. Changing the reward consumes part of the target accuracy tolerance, while restricting the policy class can lower its best attainable reward. We quantify how much statistical savings remain after accounting for both effects, and when the residual tolerance makes the guarantee uninformative (Theorem 3.6; Proposition 3.7). This puts a simpler decomposed learning problem and the original task on a common performance target.

• **Learning procedure and error propagation.** We analyze agents fitted on reference histories and then executed on generated histories. Each stage's discrepancy is weighted by the fraction of all segment rewards it can affect, yielding demonstration budgets conditional on local learning rates (Theorem 3.8; Corollary 3.1). This separates the difficulty of learning each stage from the consequences of its errors. The resulting dependence on agent count varies with the learner's rate and downstream influence; a quadratic penalty is not universal.

Concrete Gaussian examples connect these mechanisms. Stagewise and matched joint fits coincide, yet their shared error amplification can remain bounded or grow quadratically with agent count on the same dependency graph (Proposition 3.9). Comparing restricted and full feature classes then yields an exact sample-size threshold balancing estimation savings against omitted signal weighted by downstream sensitivity (Appendix J.8). With data and target matched between learners, changing coupling strength at fixed sample size and dependency graph can reverse the preferred class. Thus parameter counts and connectivity alone do not determine which model class is preferable.

Scope and limitations. This is primarily a theory paper on the statistical and learning consequences of task decomposition. Our joint PAC bounds assume conditional-value access and exact empirical maximization; the stagewise analysis uses reference demonstrations and specified learner assumptions. The PAC bounds give sufficient budgets, without characterizing minimal sample requirements, neural-optimizer convergence, or rollout complexity. The synthetic and GSM8K experiments provide qualitative context. Comparisons throughout depend on the stated objectives, policy classes, and observation resources.

2 Preliminaries and Problem Formulation

2.1 Single-Agent Reinforcement Learning (SARL) Formulation

A prompt $\boldsymbol{x} \in \mathcal{X}$ is drawn from $\mathbb{D}$. An LLM policy $\pi_{\boldsymbol{\theta}}$ generates $\boldsymbol{y} = (a_1, \ldots, a_T)$ over a finite vocabulary $\mathcal{V}$, conditioning each token on the text state $\boldsymbol{s}_t = (\boldsymbol{x}, \boldsymbol{a}_{1:t-1})$. The completed sequence receives reward $\mathrm{R}(\boldsymbol{x}, \boldsymbol{y})$. The policy class is $\Pi = \{\pi_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d\}$, where d counts free trainable coordinates (e.g., adapter parameters, not adapter rank). Define the conditional value and population objective by

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{y}\sim\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x})}[\mathrm{R}(\boldsymbol{x}, \boldsymbol{y})], \qquad \mathrm{J}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{D}}[f_{\boldsymbol{\theta}}(\boldsymbol{x})].$$

Empirical optimization and observation model. Given n independent prompts $\mathcal{S} = \{\boldsymbol{x}_j\}_{j=1}^n$ from $\mathbb{D}$, we assume access to every candidate policy's exact conditional values and use

$$\widehat{\mathrm{J}}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{j=1}^{n} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j), \qquad \widehat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta}\in\Theta} \widehat{\mathrm{J}}_n(\boldsymbol{\theta}).$$

These values depend on the candidate policy; rewards on fixed logged responses alone do not define this objective. The bounds count prompts, excluding the rollouts and computation needed to obtain conditional values or find the maximizer. They do not establish optimizer convergence. An empirical optimization error of at most τ adds τ to the regret bound.

2.2 Multi-Agent Reinforcement Learning (MARL) Formulation

We compare agents on the same task by partitioning the response into K contiguous segments. Agent i has parameters $\boldsymbol{\theta}_i \in \Theta_i \subset \mathbb{R}^{d_i}$ and observes the global prompt and preceding segments, denoted $\boldsymbol{y}^{(<i)} := \mathrm{concat}(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(i-1)})$. Generation proceeds sequentially:

$$\boldsymbol{y}^{(1)} \sim \pi_{\boldsymbol{\theta}_1}(\cdot \mid \boldsymbol{x}), \quad \boldsymbol{y}^{(i)} \sim \pi_{\boldsymbol{\theta}_i}(\cdot \mid \boldsymbol{x}, \boldsymbol{y}^{(<i)}),\ i = 2{:}K, \quad \boldsymbol{y} = \mathrm{concat}\big(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(K)}\big).$$

Writing $\overline{\boldsymbol{\theta}} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$, the joint policy is

$$\pi_{\overline{\boldsymbol{\theta}}}(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_{i=1}^{K} \pi_{\boldsymbol{\theta}_i}(\boldsymbol{y}^{(i)} \mid \boldsymbol{x}, \boldsymbol{y}^{(<i)}).$$

Sequential execution permits dependence on all earlier outputs; it does not require greedy optimization of partial rewards.

Task decomposition and optimization. For an aligned comparison, the unified reward R equals one of two normalized decompositions:

$$\overline{\mathrm{R}}_{\mathrm{dep}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{K}\sum_{i=1}^{K} \mathrm{r}_i\big(\boldsymbol{x}, \boldsymbol{y}^{(i)}, \boldsymbol{y}^{(<i)}\big), \qquad \overline{\mathrm{R}}_{\mathrm{indep}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{K}\sum_{i=1}^{K} \mathrm{r}_i\big(\boldsymbol{x}, \boldsymbol{y}^{(i)}\big). \tag{2.1}$$

Here r_i is the stage reward: the dependent form allows prefix dependence, while the independent form uses only the local segment. Reward mismatch is treated separately below. The definitions of f, J, and $\widehat{\mathrm{J}}_n$ above apply with $\boldsymbol{\theta}$ replaced by $\overline{\boldsymbol{\theta}}$ and R by the chosen $\overline{\mathrm{R}}$, using the joint policy and the same sampled prompts. The dependent result maximizes the complete empirical return over the full parameter tuple.

The independent result additionally requires history-independent policies, $\pi_{\boldsymbol{\theta}_i}(\boldsymbol{y}^{(i)} \mid \boldsymbol{x}, \boldsymbol{y}^{(<i)}) = \pi_{\boldsymbol{\theta}_i}(\boldsymbol{y}^{(i)} \mid \boldsymbol{x})$, and the Cartesian parameter space $\prod_i \Theta_i$ with disjoint trainable coordinates. With local conditional-value access, each expected stage reward depends only on $\boldsymbol{\theta}_i$, so separate empirical maximization attains the joint empirical optimum. Additive rewards alone do not suffice. A frozen shared backbone is allowed; shared trainable coordinates require a joint analysis.

2.3 Sample Complexity under the PAC Framework

For either policy class, write the preceding objectives as $\mathrm{J}(\pi)$ and $\widehat{\mathrm{J}}_n(\pi)$. We use the following probably approximately correct (PAC) criterion; Appendix C gives further background.

Definition 1 (PAC Learnability)**.** *A policy class Π is PAC-learnable if, for every $\varepsilon > 0$ and $\delta \in (0, 1)$, there is a finite $\mathrm{N}(\varepsilon, \delta)$ such that, for every $n \geq \mathrm{N}(\varepsilon, \delta)$, any empirical maximizer $\widehat{\pi} \in \arg\max_{\pi\in\Pi} \widehat{\mathrm{J}}_n(\pi)$ satisfies*

$$\Pr\left[\mathrm{J}(\widehat{\pi}) \geq \sup_{\pi\in\Pi} \mathrm{J}(\pi) - \varepsilon\right] \geq 1 - \delta,$$

where probability is over the sampled prompts.

Comparison framework. Joint PAC results measure regret relative to each stated class optimum under conditional-value access. Section 3.4 instead counts complete reference demonstrations and assumes a local learning rate; its Gaussian illustration uses expected excess MSE. Comparisons must match observation resources and performance targets and account for class-optimum gaps. Comparing sufficient budgets does not order unknown minimal sample complexities.

2.3.1 Regularization and Assumptions

All kernels and rewards are measurable, parameter sets are nonempty, and empirical maximizers are chosen measurably. Finite vocabulary and bounded horizons allow early termination to be padded with an absorbing end token. Compactness and uniform Lipschitzness below are explicit assumptions; weight decay, gradient clipping, or a soft KL penalty alone does not establish them.

Assumption 1 (Bounded Reward)**.** *The reward functions satisfy* $0 \le \mathrm{R}(\boldsymbol{x}, \boldsymbol{y}) \le 1$ *for SARL and* $0 \le \mathrm{r}_i(\boldsymbol{x}, \boldsymbol{y}) \le 1$ *for each MARL agent* i*.*

Assumption 2 (Compact Parameter Sets)**.** *The parameter spaces are compact:* $\Theta \subseteq \mathbb{R}^d$ *with* $\|\boldsymbol{\theta}\|_2 \le B$*,* $B > 0$*, for SARL, and* $\Theta_i \subseteq \mathbb{R}^{d_i}$ *with* $\|\boldsymbol{\theta}_i\|_2 \le B_i$*,* $B_i > 0$*, for each MARL agent* i*.*

Assumption 3 (Lipschitz parameterization)**.** *Uniformly over allowed prompts and histories, for SARL and for each MARL parameter block there exists a finite constant* $L_{\mathrm{step}} > 0$ *such that, for any text state* $\mathbf{s}$ *and any parameters* $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$*, the policies satisfy:*

$$\left\|\pi_{\boldsymbol{\theta}}(\cdot \mid \mathbf{s}) - \pi_{\boldsymbol{\theta}'}(\cdot \mid \mathbf{s})\right\|_1 \;\le\; L_{\mathrm{step}}\, \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2.$$

Assumption 4 (Finite Horizon)**.** *The sequence generation horizons are bounded by finite constants* $T_{\max}$ *(SARL) and* $T_{\max,i}$ *(each MARL agent* i*).*

Detailed justification and further discussion are provided in Appendix B.

3 Main Results

The results organize the decomposition question into three comparisons: how the policy class changes statistical complexity, how objective and class mismatch affect the performance target, and how local learning errors propagate under a specified training protocol. We derive sufficient prompt budgets, account for mismatch, and then study stagewise learning with separate feedback assumptions. Concrete examples illustrate what this separation reveals. Detailed proofs are provided in the Appendix.

3.1 Sample Complexity Bounds for SARL and MARL

Until Section 3.4, the results use the conditional-value observation model in Section 2.1, exact empirical maximization, and $0 < \varepsilon, \delta < 1$. We give explicit sufficient budgets U: $n \ge \lceil U \rceil$ guarantees within-class regret at most ε with probability at least $1 - \delta$. These budgets upper-bound sample complexity and are not asserted to be minimal.

We begin by establishing the PAC sample complexity bound for the SARL setting:

Theorem 3.1 (PAC sample complexity for SARL)**.** *Under Assumptions 1–4, SARL is PAC-learnable with the sufficient budget*

$$U_{\mathrm{S}}(\varepsilon, \delta) := \frac{8}{\varepsilon^2}\left[d \log\left(1 + \frac{16 L_{\mathrm{seq}} B}{\varepsilon}\right) + \log\frac{2}{\delta}\right], \qquad L_{\mathrm{seq}} = T_{\max} L_{\mathrm{step}}.$$

Theorem 3.1 relates the sufficient budget to the trainable dimension d, sequence-level regularity L_{seq}, and parameter radius B.

Next, we provide results for MARL under dependent and independent task decompositions:

Theorem 3.2 (PAC sample complexity for MARL: dependent subtasks)**.** *Consider* K *sequential, history-conditioned agents with the dependent reward in* (2.1)*. Under Assumptions 1–4, joint empirical maximization of the complete return is PAC-learnable with the sufficient budget*

$$U_{\mathrm{D}}(\varepsilon, \delta) := \frac{8}{\varepsilon^2}\left[\sum_{i=1}^{K} d_i \log\left(1 + \frac{16 K L_{\mathrm{seq},i} B_i}{\varepsilon}\right) + \log\frac{2}{\delta}\right], \quad L_{\mathrm{seq},i} = T_{\max,i} L_{\mathrm{step}}.$$

Theorem 3.3 (PAC sample complexity for MARL: independent subtasks)**.** *Consider the independent reward in* (2.1)*, history-independent segment policies, and the Cartesian trainable parameter space specified in Section 2.2. Under Assumptions 1–4, separate local empirical maximization is PAC-learnable with the sufficient budget*

$$U_{\mathrm{I}}(\varepsilon,\delta) := \frac{8}{\varepsilon^2}\left[\widetilde{d}\log\left(1+\frac{16\gamma}{\varepsilon}\right)+\log\frac{2K}{\delta}\right], \quad \widetilde{d}=\max_i d_i, \quad \gamma=\max_i L_{\mathrm{seq},i}B_i.$$

Write $\mathrm{TV}(P,Q)=\frac{1}{2}\sum_y |P(y)-Q(y)|$ for total variation. The dependent proof replaces conditional kernels one at a time, allowing each segment to depend on all earlier outputs. Allocating TV error u/K per segment gives a joint cover of radius u. This allocation enters the logarithmic covering term; concentration applies once to the complete return at accuracy ε. Thus Theorem 3.2 has no external K^2 factor or additional $\log K$ confidence term. Independent factorization instead permits simultaneous local optimization, with the largest local complexity and confidence term $\log(2K/\delta)$ governing the bound. Both results count prompts under conditional-value access. Sequential generation does not imply greedy optimization or establish a training-algorithm rate.

3.2 Relative Sample Efficiency

We compare the explicit budgets above at the same accuracy and confidence. Define

$$A = d\log(1+16L_{\mathrm{seq}}B/\varepsilon), \quad M=\widetilde{d}\log(1+16\gamma/\varepsilon), \quad S_K=\sum_i d_i\log(1+16KL_{\mathrm{seq},i}B_i/\varepsilon).$$

These comparisons concern guarantees relative to each stated class optimum. They do not establish the ordering of unknown minimal sample complexities.

Proposition 3.4 (Independent-decomposition budget comparison)**.** *Under the conditions of Theorems 3.1 and 3.3,*

$$\frac{U_{\mathrm{I}}}{U_{\mathrm{S}}}=\frac{M+\log(2K/\delta)}{A+\log(2/\delta)}.$$

Hence $U_{\mathrm{I}}\le U_{\mathrm{S}}$ whenever $M+\log K\le A$. For homogeneous segments with $d_i=d/K$ and $\gamma\le L_{\mathrm{seq}}B/K$,

$$\frac{U_{\mathrm{I}}}{U_{\mathrm{S}}}\le\frac{1}{K}+\frac{(1-1/K)\log(2/\delta)+\log K}{A+\log(2/\delta)}.$$

Proposition 3.4 makes the dimension benefit and confidence correction explicit. The displayed upper bound is approximately $1/K$ when its correction is small compared with $1/K$; independence alone does not imply this gain. A monolithic learner with the same factorized class and local-value access can exploit the same result.

Proposition 3.5 (Dependent-decomposition budget comparison)**.** *Under the conditions of Theorems 3.1 and 3.2,*

$$\frac{U_{\mathrm{D}}}{U_{\mathrm{S}}}=\frac{S_K+\log(2/\delta)}{A+\log(2/\delta)}.$$

Consequently, $U_{\mathrm{D}}\le U_{\mathrm{S}}$ if and only if $S_K\le A$. In particular, if $\sum_i d_i=d$, $L_{\mathrm{seq},i}=L_{\mathrm{seq}}/K$, and $B_i\le B$ for every i, then $U_{\mathrm{D}}\le U_{\mathrm{S}}$; equality holds when all $B_i=B$.

The comparison depends on the joint parameter-cover complexity and a common confidence term. Under the stated dimension and horizon allocation, the shorter segments offset the finer per-segment cover resolution, so dependence alone does not inflate this sufficient prompt budget. For other allocations, either budget can be smaller. These are comparisons of explicit sufficient budgets relative to their respective class optima, not an ordering of actual risks or minimal sample complexities. Specialized classes can reduce the complexity term while changing attainable task performance.

3.3 Cost of Task Misalignment

Earlier results describe how independent policy and reward decomposition can reduce sufficient learning budgets. In practice, the decomposed reward may differ from the original task reward. We now quantify this trade-off while distinguishing reward mismatch from a change in the feasible policy class.

Quantifying Task Misalignment. Let $\overline{\mathrm{R}}(\boldsymbol{x}, \boldsymbol{y})$ denote the MARL-decomposed reward and $\mathrm{R}(\boldsymbol{x}, \boldsymbol{y})$ the unified reward, both in $[0, 1]$. Define

$$\alpha := \sup_{\boldsymbol{x}, \boldsymbol{y}} \left| \overline{\mathrm{R}}(\boldsymbol{x}, \boldsymbol{y}) - \mathrm{R}(\boldsymbol{x}, \boldsymbol{y}) \right|. \tag{3.1}$$

The quantity α measures worst-case reward distortion; it does not by itself measure policy dependence or loss of expressivity.

Theorem 3.6 (PAC sample complexity for MARL under imperfect alignment)**.** *Fix a modular class Π_{M} satisfying the independent policy and parameter conditions of Theorem 3.3. Let the surrogate $\overline{\mathrm{R}}$ have the independent form in* (2.1)*. Under Assumptions 1–4, if $\varepsilon > 2\alpha$, empirical maximization of the surrogate has regret at most ε under* R*, relative to the best policy in Π_{M}, whenever*

$$n \geq \lceil U_{\mathrm{I}}(\varepsilon - 2\alpha, \delta) \rceil = \left\lceil \frac{8}{(\varepsilon - 2\alpha)^2} \left[\widetilde{d} \log\left(1 + \frac{16\gamma}{\varepsilon - 2\alpha}\right) + \log \frac{2K}{\delta} \right] \right\rceil.$$

The proof uses a 2α reward-transfer inequality on the same fixed class. For comparison with a different SARL class Π_{S}, define

$$a := \max\left\{0, \sup_{\pi \in \Pi_{\mathrm{S}}} J_R(\pi) - \sup_{\pi \in \Pi_{\mathrm{M}}} J_R(\pi)\right\}.$$

Then this sufficient guarantee requires $\varepsilon > a + 2\alpha$, and a sufficient budget is $U_{\mathrm{I}}(\varepsilon - a - 2\alpha, \delta)$. More samples reduce statistical error; they do not remove this discrepancy or class restriction. If the residual tolerance is nonpositive, the bound is uninformative, not an impossibility result.

Proposition 3.7 (Budget comparison under imperfect alignment)**.** *Let $\beta = \varepsilon - a - 2\alpha > 0$ and $M_\beta = \widetilde{d} \log(1 + 16\gamma/\beta)$. The imperfect-alignment sufficient budget is no larger than the SARL budget whenever*

$$\frac{M_\beta + \log(2K/\delta)}{A + \log(2/\delta)} \leq \left(\frac{\beta}{\varepsilon}\right)^2.$$

For $a = 0$ and negligible confidence terms, this becomes the approximate comparison

$$\kappa_d \kappa_\ell \lesssim \left(1 - \frac{2\alpha}{\varepsilon}\right)^2, \qquad \kappa_d = \frac{\widetilde{d}}{d}, \quad \kappa_\ell = \frac{\log(1 + 16\gamma/(\varepsilon - 2\alpha))}{\log(1 + 16 L_{\mathrm{seq}} B/\varepsilon)}.$$

Here, $\lesssim$ denotes an inequality up to a universal multiplicative constant.

Proposition 3.7 retains the alignment-decomposition trade-off while accounting for confidence and comparator effects. It compares sufficient budgets rather than proving a minimal-sample advantage.

3.4 Learning Errors in Stagewise Training

Theorem 3.2 allows dependent execution but assumes joint empirical maximization; it does not bound the errors produced by training agents separately. We now consider a different information model. Given n independent complete demonstrations from a fixed reference policy π^0, each agent is fitted on the reference histories preceding its segment, then frozen and executed on the preceding agents' generated outputs. One demonstration supplies K stage examples. This reference-training protocol need not optimize the complete return, and its guarantees cannot be identified with the conditional-value budgets above.

Write $H_i^0 = Y^{0,(<i)}$ for a reference history. Conditional on the training data and algorithmic randomness, define the population stage discrepancy

$$e_i(n) := \mathbb{E}_{X, H_i^0 \sim \pi^0} \, \mathrm{TV}\big(\widehat{\pi}_i(\cdot \mid X, H_i^0), \pi_i^0(\cdot \mid X, H_i^0)\big). \tag{3.2}$$

The expectation uses a fresh reference trajectory. For deterministic policies, e_i is the probability of disagreeing with the reference segment. It is not, in general, a local reward regret or an empirical training loss.

Theorem 3.8 (Propagation of stagewise learning errors)**.** *For any reference and learned sequential policies with the same normalized dependent reward in* (2.1) *and* $0 \le r_j \le 1$*,*

$$\left|J(\widehat{\pi}) - J(\pi^0)\right| \le \min\left\{1, \sum_{i=1}^{K} w_i e_i(n)\right\}, \qquad w_i := \frac{K-i+1}{K}. \tag{3.3}$$

If both policies are history-independent and the reward has the local form (2.1)*, the same statement holds with* $w_i = 1/K$*.*

The proof replaces one stage at a time, retaining reference prefixes and learned suffixes. An error at stage i can affect all $K-i+1$ remaining rewards, including through histories absent from training. Joint optimization does not remove this execution effect. The argument specializes established sequential-prediction error propagation to our normalized segment reward (Ross & Bagnell, 2010; Ross et al., 2011); its role here is to separate local learning difficulty from downstream consequences.

Corollary 3.1 (Conditional learning-rate comparison)**.** *Suppose the specified stage learners satisfy, simultaneously with probability at least* $1-\delta$*,*

$$e_i(n) \le b_i + \sqrt{C_i/n}, \qquad i = 1, \dots, K, \tag{3.4}$$

where $b_i \ge 0$ *and* $C_i \ge 0$ *are independent of* n *and include any confidence allocation. Define* $B_w = \sum_i w_i b_i$ *and* $A_w = \sum_i w_i \sqrt{C_i}$*. Then* $J(\pi^0) - J(\widehat{\pi}) \le B_w + A_w/\sqrt{n}$*, and a sufficient demonstration budget for error* $\varepsilon > B_w$ *is*

$$V_w(\varepsilon, \delta) := \frac{A_w^2}{(\varepsilon - B_w)^2}, \qquad n \ge \max\{1, \lceil V_w \rceil\}. \tag{3.5}$$

For a target optimum $J^\star$*, add* $g_0 = \max\{0, J^\star - J(\pi^0)\}$ *to the error bound and replace the residual tolerance by* $\varepsilon - g_0 - B_w > 0$*.*

The premise (3.4) must be established for the training algorithm and data; it is not implied by small local reward losses or by running Adam. For $b_i = 0$ and $C_i = C_0$, the dependent and independent certificates are respectively $(K+1)^2 C_0/(4\varepsilon^2)$ and C_0/ε^2. A justified faster rate $e_i \le C_i/n$ instead gives $V_w = \sum_i w_i C_i/\varepsilon$. Appendix J.5 derives this faster rate for a concrete realizable finite-class learner and also gives a complementary product-class bound. Thus a quadratic factor is not an unavoidable consequence of stagewise training.

For $b_i = 0$ and $C_{\mathrm{S}} > 0$, consider a SARL learner with its own proved, zero-floor guarantee $J(\pi^0) - J(\widehat{\pi}_{\mathrm{S}}) \le \sqrt{C_{\mathrm{S}}/n}$ under matched feedback and the same reward and reference. The dependent-to-SARL certificate ratio is A_w^2/C_{S}. If additionally $C_i = C_{\mathrm{S}}/K$, it is $(K+1)^2/(4K)$, not K^2. These relations compare sufficient algorithm-specific budgets, not actual risks; parameter counts alone do not establish the assumed C_i. Appendix J gives proofs, nonzero-floor comparisons, and a joint estimation/optimization error decomposition.

Illustrating the framework with an exact comparison. To separate learning error from task sensitivity, we compare specified stagewise and joint estimators using normalized expected excess MSE: $\mathcal{E}_K := K^{-1}\mathbb{E}\|\widehat{m}(X) - \mathbb{E}[Y \mid X]\|_2^2$. The outer expectation includes training data and a fresh test input. This criterion is distinct from bounded-reward PAC regret.

Proposition 3.9 (Exact learning and propagation with a matched joint learner)**.** *Let* $Y = CY + f(X) + \xi$*, where* C *is known and strictly lower triangular,* $f_i(X) = w_i^\top x_i$*,* $x_i \sim \mathcal{N}(0, I_p)$*, and* $\xi_i \sim \mathcal{N}(0, \sigma^2)$*, with* $p \ge 1$*, known* $\sigma^2 > 0$*, and all inputs and innovations mutually independent. Given* $n > p+1$ *independent complete observations* (X, Y)*, fit each* w_i *by ordinary least squares on* $Z_i = Y_i - \sum_{j<i} C_{ij} Y_j$*, then execute recursively on predicted histories. Write* $G = (I-C)^{-1}$ *and* $\mathcal{A}_K(C) = \|G\|_F^2/K$*. This stagewise learner and joint Gaussian trajectory maximum likelihood over the same product linear class give the same estimator, with normalized expected excess MSE*

$$\mathcal{E}_K^{\mathrm{stage}} = \mathcal{E}_K^{\mathrm{joint}} = \frac{\sigma^2 p}{n-p-1}\,\mathcal{A}_K(C). \tag{3.6}$$

For $C_{ij}^{(a)} = a/(i-1)$ *when* $i > j$*, and zero otherwise, write* $\mathcal{A}_{K,a} := \mathcal{A}_K(C^{(a)})$*. Then*

$$\mathcal{A}_{K,1} = 1 + \frac{1}{K}\sum_{j=1}^{K-1}\frac{K-j}{j^2} < 1 + \frac{\pi^2}{6}, \quad \mathcal{A}_{K,2} = 1 + \frac{4}{K}\sum_{j=1}^{K-1}\frac{\sum_{i=j+1}^{K} i^2}{j^2(j+1)^2} = \Theta(K^2). \tag{3.7}$$

Appendix J.7 gives the proof and exact budgets. Both couplings share the complete dependency graph, local estimation problems, and trainable dimension Kp. At fixed p, σ^2, and target absolute excess MSE, their budgets are respectively bounded in K and quadratic in K: connectivity alone does not determine propagation. Joint likelihood fitting shares both effects and does not eliminate execution sensitivity. Bayes MSE is also amplified by $\mathcal{A}_K(C)$, giving the common excess-to-Bayes-MSE ratio $p/(n-p-1)$. This identifies shared task sensitivity, without predicting architectural differences or neural training curves. Appendix J.8 derives an exact sample-size threshold for beneficial feature restriction and shows that coupling strength can reverse the preferred class.

3.5 Empirical Study

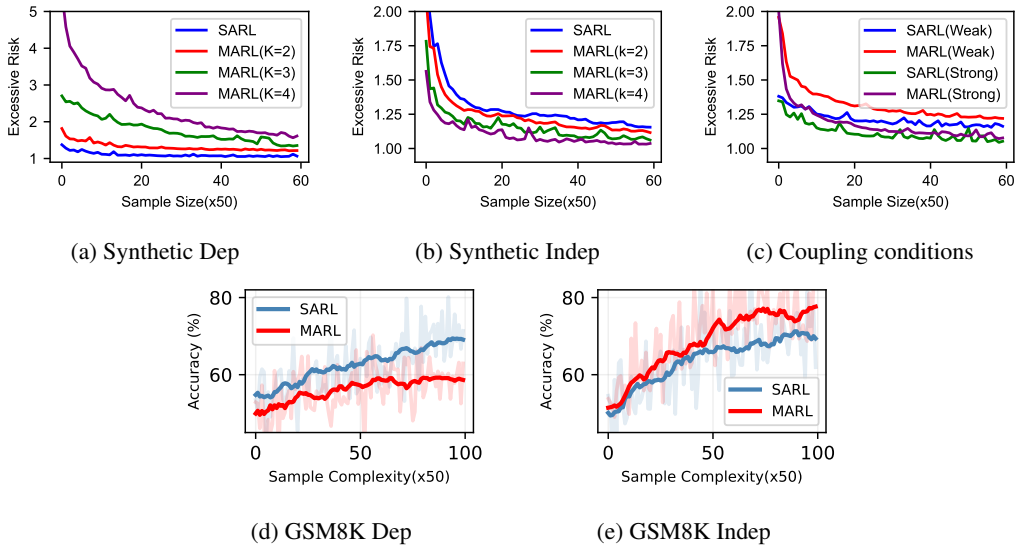
The study examines whether specialization benefits independent subtasks and whether downstream sensitivity can offset that benefit when subtasks interact. GSM8K provides a small-scale comparison across independent and dependent task constructions.

Experimental Setup. The synthetic regression task has independent and dependent variants:

$$y_i = \begin{cases} \mathbf{w}_i^\top \mathbf{x}_i + \xi_i, & \text{(independent subtasks)}, \\ \mathbf{w}_i^\top \mathbf{x}_i + \lambda \cdot \text{average}(y^{(<i)}) + \xi_i, & \text{(dependent subtasks)}, \end{cases}$$

where local features are Gaussian and $\xi_i \sim \mathcal{N}(0,1)$. SARL learns the complete output with one model; MARL uses sequentially trained specialized agents with distinct parameters. The experiments use one-layer MLPs for synthetic tasks and `Qwen2.5-1.5B-Instruct` for GSM8K (Cobbe et al., 2021). Independent GSM8K tasks concatenate two unrelated questions; dependent tasks split a solution into thinking and answering. Curves average five trials. Appendix H gives reported settings and proposed, unexecuted follow-up protocols.

Results and interpretation. Figure 1 shows lower reported risk for MARL on independent tasks and for SARL on dependent tasks, with larger gaps as K increases. These trends are consistent with specialization simplifying local learning while downstream errors can offset that benefit under the studied procedures. Panel (c) was designed to compare $\lambda = 0.1$ (Strong) with $\lambda = 1$ (Weak); the smaller Strong-condition gap is consistent with weaker interactions making local decomposition more effective. Here λ controls coupling; relating it to α requires specifying the local and evaluation



(a) Synthetic Dep (b) Synthetic Indep (c) Coupling conditions

(d) GSM8K Dep (e) GSM8K Indep

Figure 1: SARL–MARL comparisons for (a) dependent and (b) independent synthetic tasks, (c) lower/higher coupling (Strong/Weak), and (d) dependent and (e) independent GSM8K constructions. The Strong/Weak labels describe intended compatibility with local decomposition, not measured reward discrepancy α.

rewards. The GSM8K experiment with Qwen2.5-1.5B-Instruct show the corresponding accuracy ordering later in training, supporting the relevance of task structure.

4 Related Work

Table 1 summarizes related analyses. Hierarchical RL already compares hierarchical and monolithic learning through state, action, and temporal abstraction (Robert et al., 2023). Our contribution combines class complexity, reward/class mismatch, and stagewise propagation in SARL–MARL comparisons. The conditional-value and reference-demonstration analyses isolate policy estimation and learning errors; they do not model exploration or learning the hierarchy. Appendix K gives the broader comparison.

Table 1: Selected theoretical comparisons; observation models and performance comparators differ.

Work	**Setting**	**Central result**	**Comparison focus**
Robert et al. (2023)	Goal-conditioned HRL; interaction episodes	PAC lower bound for a learned hierarchy	Hierarchical vs. monolithic learning
Chen et al. (2021)	Factored MDPs; environment interaction	Regret upper and lower bounds	Transition/reward structure vs. tabular learning
Wang et al. (2021)	Cooperative multi-agent fitted Q-iteration	Value class and data distribution affect convergence	Value factorization and credit assignment
Ross & Bagnell (2010); Ross et al. (2011)	Imitation learning; expert feedback	Compounding errors and interactive reductions	Reference vs. induced histories
Ours	Sequential policies; conditional values or demonstrations	Sufficient budgets, reward/class gaps, and propagation	**SARL–MARL comparison across three mechanisms**

5 Concluding Discussion

Our analysis shows that task decomposition alone does not determine SARL–MARL sample efficiency. The comparison depends on class complexity, reward and policy-class mismatch, and error propagation. Under conditional-value access, independent factorization can reduce sufficient learning budgets, while fully dependent modular policies can match the single-agent certificate under matched complexity. For stagewise training, dependence matters through local learning rates and downstream error amplification, not agent count alone. The exact Gaussian analysis further reveals an estimation–approximation trade-off: feature restriction reduces estimation error but incurs omitted-signal error weighted by downstream sensitivity. Thus, even with the same dependency graph, feature classes, and sample size, coupling strength can reverse which class has lower expected excess risk (Appendix J.8). Meaningful SARL–MARL comparisons therefore require matched resources and targets, together with explicit accounting for class complexity, attainable performance, training procedure, and downstream sensitivity.

Limitations and Future Work. The joint PAC guarantees assume independent prompts, conditional-value access, and empirical maximization, and therefore characterize sufficient prompt budgets rather than minimal sample requirements, rollout complexity, or neural-optimizer convergence. The stagewise results use reference demonstrations and require separately justified local learning rates. The Gaussian analysis assumes known coupling, independent Gaussian local designs and innovations, and exact least-squares estimation. Extending the framework to unknown or learned coupling, shared trainable parameters, approximate optimization, and fully interactive data collection—with explicit exploration and observation models—is an important direction for future work.

AI USE STATEMENT

Generative AI tools assisted with drafting and polishing portions of the manuscript and preparing verification code. The authors remain responsible for all claims and proofs.

REFERENCES

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020. ISSN 2577-0187. doi: 10.1137/20m1336072. URL http://dx.doi.org/10.1137/20M1336072.

Weize Chen, Jiarui Yuan, Chen Qian, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Optima: Optimizing effectiveness and efficiency for LLM-based multi-agent system. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11534–11557, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.601. URL https://aclanthology.org/2025.findings-acl.601/.

Xiaoyu Chen, Jiachen Hu, Lihong Li, and Liwei Wang. Efficient reinforcement learning in factored MDPs with application to constrained RL, 2021. URL https://arxiv.org/abs/2008.13319.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

Michael Kearns and Daphne Koller. Efficient reinforcement learning in factored mdps. In *IJCAI*, volume 16, pp. 740–747, 1999.

Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A. Ali Heydari, Yao Yan, Zhihan Zhang, Yuchen Zhuang, Yun Liu, Mark Malhotra, Paul Pu Liang, Hae Won Park, Yuzhe Yang, Xuhai Xu, Yilun Du, Shwetak Patel, Tim Althoff, Daniel McDuff, and Xin Liu. Towards a science of scaling agent systems, 2026. URL https://arxiv.org/abs/2512.08296.

Daphne Koller and Ronald Parr. Computing factored value functions for policies in structured mdps. In *IJCAI*, volume 99, pp. 1332–1339, 1999.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning, 2018. URL https://arxiv.org/abs/1803.11485.

Arnaud Robert, Ciara Pike-Burke, and Aldo Faisal. Sample complexity of goal-conditioned hierarchical reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 62696–62712. Curran Associates, Inc., 2023. doi: 10.52202/075280-2738.

Stephane Ross and Drew Bagnell. Efficient reductions for imitation learning. In Yee Whye Teh and Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 661–668, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/ross10a.html.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/ross11a.html.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning, 2017. URL https://arxiv.org/abs/1706.05296.

Dat Tran and Douwe Kiela. Single-agent LLMs outperform multi-agent systems on multi-hop reasoning under equal thinking token budgets, 2026. URL https://arxiv.org/abs/2604.02460.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*, 2025.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Jianhao Wang, Zhizhou Ren, Beining Han, Jianing Ye, and Chongjie Zhang. Towards understanding cooperative multi-agent q-learning with value factorization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29142–29155. Curran Associates, Inc., 2021.

A RL TRAINING FOR LLMS

Reinforcement learning (RL) trains and fine-tunes large language models (LLMs) using feedback on generated responses. This feedback can reward accuracy, contextual coherence, and alignment with human preferences, allowing policies to improve through interaction.

A.1 COMMON RL OBJECTIVES FOR LLMS

A common objective is to maximize expected sequence-level reward. Given a prompt $\boldsymbol{x} \sim \mathcal{D}$ and a token sequence $\boldsymbol{y} = (a_1, \ldots, a_T)$ drawn from the policy $\pi_{\boldsymbol{\theta}}$, this objective is

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D},\, \boldsymbol{y}\sim\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x})}[R(\boldsymbol{x}, \boldsymbol{y})], \tag{A.1}$$

where $R(\boldsymbol{x}, \boldsymbol{y}) \in [0, 1]$ denotes a scalar reward evaluated after the sequence is produced.

Common reward sources include:

- **Human preference scores**: Human annotators evaluate generated outputs for factors like coherence, accuracy, and relevance.
- **Task-specific correctness indicators**: Binary or continuous evaluations of task performance such as factual correctness or logical accuracy.
- **Automated evaluation metrics**: Metrics such as BLEU, ROUGE, and BERTScore measure linguistic quality, coherence, and semantic similarity to reference texts.

A.2 OPTIMIZATION ALGORITHMS

Policy-gradient methods for LLM training often use regularization or trust-region constraints to control policy updates. Examples include Proximal Policy Optimization (PPO) and Group Relative Policy Optimization (GRPO):

Proximal Policy Optimization (PPO) PPO optimizes a clipped surrogate objective to stabilize updates:

$$\begin{aligned} L^{\mathrm{PPO}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim\pi_{\boldsymbol{\theta}^-}}\Big[\min\big(r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{y})A^-,\ \mathrm{clip}(r_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{y}), 1-\epsilon, 1+\epsilon)A^-\big)\Big] \\ - \beta\, \mathbb{E}_{\boldsymbol{x}}[D_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}}\|\pi_{\boldsymbol{\theta}^-})], \end{aligned} \tag{A.2}$$

where:

- $r_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\boldsymbol{\theta}^-}(\boldsymbol{y}|\boldsymbol{x})}$ denotes the importance ratio relative to a previous policy $\pi_{\boldsymbol{\theta}^-}$.
- $A^-(\boldsymbol{x}, \boldsymbol{y}) = R(\boldsymbol{x}, \boldsymbol{y}) - V^-(\boldsymbol{x})$ is the advantage estimate.
- $\epsilon > 0$ controls the update step size, and $\beta \geq 0$ regularizes updates via KL divergence constraints.

This method is widely applied due to its stability and efficiency in LLM fine-tuning.

Groupwise Relative Policy Optimization (GRPO) GRPO uses group-based reward comparisons to enhance training stability and reduce variance:

$$\begin{aligned} L^{\mathrm{GRPO}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D},\, \boldsymbol{y}^{1:G}\sim(\pi_{\boldsymbol{\theta}^-})^{\otimes G}}\left[\frac{1}{G}\sum_{g=1}^{G} w(\boldsymbol{x}, \boldsymbol{y}^{1:G}, g)\, r_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}^{(g)})\right] \\ - \beta\, \mathbb{E}_{\boldsymbol{x}}[D_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}}\|\pi_{\boldsymbol{\theta}^-})], \end{aligned} \tag{A.3}$$

where:

- $w(\boldsymbol{x}, \boldsymbol{y}^{1:G}, g)$ represents a group-relative reward, typically constructed to be centered, reducing estimator variance.
- G denotes the number of completions sampled per prompt, enabling inter-group reward comparison.

GRPO improves training efficiency and reliability, especially when evaluating subtler quality differences between multiple generated sequences.

Common features. These training frameworks combine:

1. Optimization of a sequence-level reward function reflecting task-specific performance, correctness, or human preferences.
2. Autoregressive policies generating sequences token-by-token, enabling sequential decision-making.
3. KL-regularized objectives or trust-region constraints ensuring stable policy updates.
4. Policy-gradient methods such as PPO and GRPO for efficient, stable policy optimization.

We study the sequence-level reward objective under an idealized statistical model. The guarantees concern exact empirical maximizers with access to conditional expected rewards; they do not establish convergence or rollout complexity for PPO, GRPO, or other practical optimizers.

B Analytical Setting

B.1 LLM RL in Practice

Reinforcement learning from human feedback (RLHF) fine-tunes LLMs using rewards derived from human preference comparisons or proxy reward models. The model acts as an agent that receives prompts, generates responses, and obtains sequence-level feedback, adapting its behavior through interaction.

Multi-agent RL (MARL) decomposes the overall task into subtasks handled by specialized agents, such as retrieval, planning, generation, and verification. Agents interact sequentially or in parallel and receive coordinated feedback on their contributions. Such architectures support specialization and parallelism in complex LLM applications.

Our PAC analysis compares a single policy generating the complete sequence with specialized policies generating successive segments, as in a planner–solver–verifier pipeline. Both receive sequence-level rewards. The sufficient prompt budgets use a common observation model: for each sampled prompt x, the conditional value $f_\theta(x) = \mathbb{E}_{y\sim\pi_\theta(\cdot|x)}R(x, y)$ is available exactly for every candidate policy. Thus n counts independent prompts, not sampled trajectories or reward queries. Computing these values and finding a global empirical maximizer are separate requirements. Even under a shared evaluation reward, different policy classes can have different attainable optima; comparison to a SARL optimum therefore includes the class gap specified in the alignment analysis.

B.2 Regularity Assumptions

Training practices motivate the following assumptions, which must be imposed or verified for the policy classes under study:

- Bounded rewards. In RLHF and related paradigms, rewards are typically normalized or clipped to a bounded interval, commonly $[0, 1]$. This reflects both the probabilistic nature of preference-based rewards and the need for stable optimization.
- Compact parameter space. Practical fine-tuning of LLMs employs regularization strategies such as weight decay, KL penalties around a reference policy, or low-rank adapter updates. These motivate restricted parameter classes, but do not by themselves make the feasible domain compact. Our analysis explicitly assumes a closed, norm-bounded feasible parameter set.
- Lipschitz parameterization. We assume a finite parameter Lipschitz constant uniformly over the prompts and histories considered. Smooth neural parameterizations can support this assumption under suitable boundedness conditions. Gradient clipping controls update gradients and does not itself prove uniform sensitivity bounds for action distributions.
- Finite horizon. A hard cap on generation length supplies the finite horizon assumed for the policies considered here.

These assumptions yield finite covers of the induced conditional-value class. Exact conditional-value access is a separate idealization. Applying the resulting statistical guarantees to practical RL training also requires accounting for value estimation and optimization.

C THE PAC FRAMEWORK AND SAMPLE COMPLEXITY ANALYSIS

The *Probably Approximately Correct* (PAC) learning framework, introduced by Valiant (Valiant, 1984), relates sample size to population performance and confidence. Let $\mathcal{D}$ be an unknown distribution over input space $\mathcal{X}$, and let $\mathcal{H}$ denote a hypothesis class mapping $\mathcal{X}$ to labels or real-valued outputs. Given i.i.d. samples $S = \{x_i\}_{i=1}^n \sim \mathcal{D}^n$, the goal is to identify a hypothesis $h \in \mathcal{H}$ whose population risk is close to the optimum achievable within $\mathcal{H}$.

PAC Learnability. A hypothesis class $\mathcal{H}$ is said to be PAC-learnable if there exists a function $N(\varepsilon, \delta, \mathcal{H})$ such that for all $\varepsilon, \delta \in (0,1)$, with at least $n \geq N(\varepsilon, \delta, \mathcal{H})$ samples, the empirical risk minimizer $\widehat{h}$ satisfies

$$\Pr\left(L(\widehat{h}) \leq \inf_{h \in \mathcal{H}} L(h) + \varepsilon\right) \geq 1 - \delta,$$

where $L(h) := \mathbb{E}_{x \sim \mathcal{D}}[\ell(h(x))]$ denotes the population risk under a loss ℓ. Here ε quantifies the approximation tolerance and δ the confidence level.

Sample Complexity. A sufficient threshold $N(\varepsilon, \delta, \mathcal{H})$ gives an upper bound on the *sample complexity*. In supervised learning, sample complexity is characterized via combinatorial measures such as VC-dimension or via capacity measures such as covering numbers and Rademacher complexity. The sample unit depends on the observation model. Here it is an independent prompt, with exact conditional expected rewards available for all candidate policies; our bounds do not count environment interactions, rollouts, or reward queries needed to obtain those values.

Application to LLM–RL. We apply the PAC framework to single-agent RL (SARL) and multi-agent RL (MARL). Policies $\pi \in \Pi$ map text states to action distributions over tokens, and the objective is to maximize the sequence-level reward

$$J(\pi) \;=\; \mathbb{E}_{x \sim \mathcal{D},\, y \sim \pi(\cdot|x)}[R(x,y)].$$

The PAC guarantee quantifies how many prompts suffice for the exact empirical maximizer $\widehat{\pi}$ to achieve $J(\widehat{\pi}) \geq \sup_{\pi \in \Pi} J(\pi) - \varepsilon$ with probability at least $1 - \delta$. The explicit sufficient budgets U_{S}, U_{D}, and U_{I} apply to the stated SARL, dependent MARL, and independent MARL classes. Comparing these budgets does not compare minimum sample complexities. Each guarantee uses its own within-class optimum; comparison across classes additionally accounts for the attainable-value gap. PAC generalization alone does not establish optimizer convergence.

Role of Regularity Assumptions. Finite, non-asymptotic sample complexity bounds require control of the effective complexity of the policy class Π. We assume bounded rewards, compact parameter spaces, uniform Lipschitz parameterizations, and finite horizons. These assumptions imply finite covering numbers for the induced function class $\{f_\theta(x) := \mathbb{E}_{y \sim \pi_\theta(\cdot|x)}[R(x,y)]\}$ and yield PAC guarantees in the stated observation model. As discussed in Appendix B, practical regularization and gradient clipping do not automatically verify these assumptions.

Interpretation. Our sufficient prompt budgets identify conditions under which the independent MARL bound is smaller than the SARL bound. They do not establish a separation between minimum sample complexities, nor an intrinsic K^2 penalty for dependent MARL. Their implications for agentic LLM design are conditional on the stated observation model, policy classes, and exact empirical optimization.

D PAC SAMPLE COMPLEXITY FOR SARL

The SARL PAC bound follows from a covering argument for the conditional-value class.

The population objective is

$$\mathrm{J}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{D},\, \boldsymbol{y}\sim\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x})}\big[\mathrm{R}(\boldsymbol{x},\boldsymbol{y})\big],$$

In the conditional-value access model, the samples are iid prompts $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^n$. Assumption 1 ensures that the conditional value is bounded:

$$f_{\boldsymbol{\theta}}(\mathbf{x}) \;:=\; \mathbb{E}_{y\sim\boldsymbol{\pi}_{\boldsymbol{\theta}}(\cdot|\mathbf{x})}[\mathrm{R}(\mathbf{x},\mathbf{y})] \in [0,1], \qquad \mathrm{J}(\boldsymbol{\theta}) \;:= \mathbb{E}_{\mathbf{x}\sim\mathbb{D}}\big[f_{\boldsymbol{\theta}}(\mathbf{x})\big],$$

The corresponding value class is $\mathcal{F} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\} \subset [0,1]^{\mathcal{X}}$. For prompts $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathbb{D}$, the empirical objective is

$$\widehat{\mathrm{J}}_n(\boldsymbol{\theta}) \;:=\; \frac{1}{n}\sum_{i=1}^{n} f_{\boldsymbol{\theta}}(\mathbf{x}_i).$$

The learner maximizes this objective using the conditional expectations for each candidate policy. Thus n counts prompts. Approximate optimization satisfying $\widehat{\mathrm{J}}_n(\widehat{\boldsymbol{\theta}}) \geq \sup_{\boldsymbol{\theta}} \widehat{\mathrm{J}}_n(\boldsymbol{\theta}) - \tau$ adds the tolerance $\tau \geq 0$ to the regret guarantee.

D.1 From Per-Step Control to Sequence Distributions

The total variation (TV) distance between probability distributions $\mathbb{P}$ and $\mathbb{Q}$ is

$$\mathrm{TV}(\mathbb{P} \mid \mathbb{Q}) = \frac{1}{2}\sum_{x\in\mathcal{X}} |\mathbb{P}(x) - \mathbb{Q}(x)|.$$

Lemma D.1 (TV control for autoregressive sequence)**.** *Under Assumption 3 and 4, let* $L_{\mathrm{seq}} := T_{\max} L_{\mathrm{step}}$*. Then, for any fixed* $\mathbf{x}$ *and the induced full-sequence laws* $\mathbb{P}_{\boldsymbol{\theta}}(\cdot \mid \mathbf{x})$ *and* $\mathbb{P}_{\boldsymbol{\theta}'}(\cdot \mid \mathbf{x})$*, we have*

$$\mathrm{TV}(\mathbb{P}_{\boldsymbol{\theta}}(\cdot \mid \mathbf{x}),\, \mathbb{P}_{\boldsymbol{\theta}'}(\cdot \mid \mathbf{x})) \;\leq\; L_{\mathrm{seq}}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2.$$

Proof. Pad early termination with an absorbing end token. Replace the token kernels one at a time to form hybrid sequence laws. Adjacent hybrids have the same entering-history law, and applying their identical downstream kernels cannot increase total variation. The triangle inequality therefore gives

$$\mathrm{TV}(\mathbb{P}_{\boldsymbol{\theta}}(\cdot \mid \mathbf{x}),\, \mathbb{P}_{\boldsymbol{\theta}'}(\cdot \mid \mathbf{x})) \leq \sum_{t=1}^{T_{\max}} \sup_{\mathbf{s}_t} \frac{1}{2}\big\|\boldsymbol{\pi}_{\boldsymbol{\theta}}(\cdot \mid \mathbf{s}_t) - \boldsymbol{\pi}_{\boldsymbol{\theta}'}(\cdot \mid \mathbf{s}_t)\big\|_1$$

Assumption 3 gives

$$\sum_{t=1}^{T_{\max}} \sup_{\mathbf{s}_t} \frac{1}{2}\big\|\boldsymbol{\pi}_{\boldsymbol{\theta}}(\cdot \mid \mathbf{s}_t) - \boldsymbol{\pi}_{\boldsymbol{\theta}'}(\cdot \mid \mathbf{s}_t)\big\|_1 \leq T_{\max} L_{\mathrm{step}}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2.$$

Substituting $L_{\mathrm{seq}} = T_{\max} L_{\mathrm{step}}$ yields

$$\mathrm{TV}(\mathbb{P}_{\boldsymbol{\theta}}(\cdot \mid \mathbf{x}),\, \mathbb{P}_{\boldsymbol{\theta}'}(\cdot \mid \mathbf{x})) \leq L_{\mathrm{seq}}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$$

□

Lemma D.2 (Parameter-to-value Lipschitzness)**.** *For any* $\mathbf{x}$*,*

$$|f_{\boldsymbol{\theta}}(\mathbf{x}) - f_{\boldsymbol{\theta}'}(\mathbf{x})| \;\leq\; L_{\mathrm{seq}}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2.$$

Proof. The definition of $f_{\boldsymbol{\theta}}(\cdot)$ and Lemma D.1 give

$$\begin{aligned} |f_{\boldsymbol{\theta}}(\mathbf{x}) - f_{\boldsymbol{\theta}'}(\mathbf{x})| &= \Big|\mathbb{E}_{\mathbf{y}\sim\mathbb{P}_{\boldsymbol{\theta}}}[\mathrm{R}(\mathbf{x},\mathbf{y})] - \mathbb{E}_{\mathbf{y}\sim\mathbb{P}_{\boldsymbol{\theta}'}}[\mathrm{R}(\mathbf{x},\mathbf{y})]\Big|, \\ &\leq \mathrm{TV}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}_{\boldsymbol{\theta}'}), \\ &\leq L_{\mathrm{seq}}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 \end{aligned}$$

□

Lemma D.3 (Parameter-to-function cover). *Let $\Theta \subset \mathbb{R}^d$ be a compact parameter set and let $\{f_\theta : \mathcal{X} \to [0,1]\}_{\theta\in\Theta}$ be a function class. Assume the parameterization is Lipschitz in the uniform (sup) norm: there exists $L > 0$ such that for all $\theta, \theta' \in \Theta$,*

$$\|f_\theta - f_{\theta'}\|_\infty \;\le\; L\,\|\theta - \theta'\|_2.$$

Then for every $\varepsilon > 0$,

$$\mathcal{N}(\varepsilon,\, \{f_\theta : \theta \in \Theta\},\, \|\cdot\|_\infty) \;\le\; \mathcal{N}\big(\tfrac{\varepsilon}{L},\, \Theta,\, \|\cdot\|_2\big)\,.$$

Proof. Let $\{u_1, \ldots, u_M\}$ be a (ε/L)-cover of Θ under $\|\cdot\|_2$, i.e., $M = \mathcal{N}(\varepsilon/L, \Theta, \|\cdot\|_2)$ and for every $\theta \in \Theta$ there exists u_j with $\|\theta - u_j\|_2 \le \varepsilon/L$. By the assumed uniform Lipschitz inequality,

$$\|f_\theta - f_{u_j}\|_\infty \;\le\; L\,\|\theta - u_j\|_2 \;\le\; \varepsilon,$$

so $\{f_{u_1}, \ldots, f_{u_M}\}$ is an ε-cover of the function class in $\|\cdot\|_\infty$. □

D.2 Sample Complexity via Covering the Value Class

Lemma D.4 (Sample Complexity via Covering the Value Class). *Let $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty)$ denote the covering number of a function class $\mathcal{F}$ with respect to the ℓ_∞-norm, and let $H_{\mathcal{F}}(\varepsilon) := \log \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty)$. Then, under SARL setting and Assumption 2, the following holds:*

$$H_{\mathcal{F}}(\varepsilon) \le d \log\left(1 + \frac{2L_{\text{seq}}B}{\varepsilon}\right).$$

Proof. For $L_{\text{seq}} > 0$, Lemma D.3 relates the covering numbers of the function class $\mathcal{F}$ and the parameter space Θ:

$$\mathcal{N}\left(\varepsilon, \mathcal{F}, \|\cdot\|_\infty\right) \le \mathcal{N}\left(\frac{\varepsilon}{L_{\text{seq}}}, \Theta, \|\cdot\|_2\right).$$

An ℓ_2-cover of the parameter space induces a cover of $\mathcal{F}$ in the ℓ_∞-norm. The required resolution in Θ is scaled by L_{seq}, the parameter-to-value Lipschitz constant.

Since the parameter set lies in an ℓ_2-ball of radius B in $\mathbb{R}^d$, the Euclidean covering bound gives

$$\log \mathcal{N}\left(\eta, \Theta, \|\cdot\|_2\right) \le d \log\left(1 + \frac{2B}{\eta}\right),$$

for every $\eta > 0$. A maximal η-separated subset of Θ is an η-net with centers in Θ; disjoint balls of radius $\eta/2$ around its centers lie in a ball of radius $B + \eta/2$, which proves the cardinality bound.

Combining the two bounds gives the entropy bound for $\mathcal{F}$:

$$H_{\mathcal{F}}(\varepsilon) \le d \log\left(1 + \frac{2L_{\text{seq}}B}{\varepsilon}\right).$$

If $L_{\text{seq}} = 0$, the value class is a singleton and the same bound holds with zero entropy. □

D.3 Proof of SARL Sample Complexity

The value-class entropy bound yields Theorem 3.1 through uniform convergence.

Proof. Let $\mathcal{C} \subset \mathcal{F}$ be a sup-norm $\varepsilon/8$-net, and put $M = |\mathcal{C}|$. For each fixed $g \in \mathcal{C}$, Hoeffding's inequality gives

$$\Pr\left(\left|\frac{1}{n}\sum_{j=1}^{n} g(\mathbf{x}_j) - \mathbb{E}g(\mathbf{x})\right| > \frac{\varepsilon}{4}\right) \le 2\exp\left(-\frac{n\varepsilon^2}{8}\right).$$

A union bound over the net shows that all these deviations are at most $\varepsilon/4$ except on an event of probability at most $2M\exp\big(-n\varepsilon^2/8\big)$. For every $f \in \mathcal{F}$, choose a net representative g with

$\|f-g\|_\infty \le \varepsilon/8$. On the complementary event,

$$\left|\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x}_j) - \mathbb{E}f(\mathbf{x})\right| \le \left|\frac{1}{n}\sum_{j=1}^{n}(f-g)(\mathbf{x}_j)\right| + \left|\frac{1}{n}\sum_{j=1}^{n} g(\mathbf{x}_j) - \mathbb{E}g(\mathbf{x})\right| + |\mathbb{E}(g-f)(\mathbf{x})|$$
$$\le \frac{\varepsilon}{8} + \frac{\varepsilon}{4} + \frac{\varepsilon}{8} = \frac{\varepsilon}{2}.$$

Consequently,

$$\sup_{\boldsymbol{\theta}\in\Theta} |\widehat{\mathrm{J}}_n(\boldsymbol{\theta}) - \mathrm{J}(\boldsymbol{\theta})| \le \varepsilon/2.$$

The compact Lipschitz parameterization has a countable dense subset, so the displayed supremum is measurable whenever the individual value functions are measurable.

The entropy bound above gives

$$\log M \le d \log\left(1 + \frac{16 L_{\mathrm{seq}} B}{\varepsilon}\right).$$

Thus it suffices to choose any integer $n \ge \lceil U_{\mathrm{S}}(\varepsilon,\delta)\rceil$, where

$$U_{\mathrm{S}}(\varepsilon,\delta) := \frac{8}{\varepsilon^2}\left[d \log\left(1 + \frac{16 L_{\mathrm{seq}} B}{\varepsilon}\right) + \log\frac{2}{\delta}\right].$$

Finally, for an empirical maximizer with tolerance $\tau \ge 0$,

$$\sup_{\boldsymbol{\theta}} \mathrm{J}(\boldsymbol{\theta}) \le \sup_{\boldsymbol{\theta}} \widehat{\mathrm{J}}_n(\boldsymbol{\theta}) + \varepsilon/2$$
$$\le \widehat{\mathrm{J}}_n(\widehat{\boldsymbol{\theta}}) + \tau + \varepsilon/2 \le \mathrm{J}(\widehat{\boldsymbol{\theta}}) + \tau + \varepsilon.$$

Taking $\tau = 0$ proves the result for exact empirical maximization. The budget counts iid prompts under conditional-value access. □

E Sample Complexity for MARL

We derive uniform-convergence PAC bounds for MARL under *perfect task decomposition*. Throughout, rewards are bounded in $[0,1]$ and expectations are taken over the data distribution $\mathcal{D}$ and the policies' internal randomness.

A common sequence-level reward. The comparison with SARL uses the *same* sequence-level reward $\mathrm{R}(\boldsymbol{x},\boldsymbol{y})$, expressed in one of two normalized additive forms:

$$\overline{\mathrm{R}}_{\mathrm{dep}}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{K}\sum_{i=1}^{K} \mathrm{r}_i\big(\boldsymbol{x}, \boldsymbol{y}^{(i)}, \boldsymbol{y}^{(<i)}\big), \tag{E.1}$$

$$\overline{\mathrm{R}}_{\mathrm{indep}}(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{K}\sum_{i=1}^{K} \mathrm{r}_i\big(\boldsymbol{x}, \boldsymbol{y}^{(i)}\big). \tag{E.2}$$

As in (2.1), dependent rewards may use the preceding segments $\boldsymbol{y}^{(<i)}$. The independent reward is a sum of local terms; the corresponding guarantee also requires conditionally independent policies and a Cartesian product parameter set.

The MARL population objective is

$$\mathrm{J}_{\mathrm{MARL}}(\boldsymbol{\pi}_{\boldsymbol{\theta}_{\mathrm{MARL}}}) = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{D},\, \boldsymbol{y}\sim\boldsymbol{\pi}_{\boldsymbol{\theta}_{\mathrm{MARL}}}(\cdot|\boldsymbol{x})}\big[\overline{\mathrm{R}}_{\mathrm{dep/indep}}(\mathbf{x},\mathbf{y})\big]$$

Empirical optimization. The learner uses iid prompts $\mathcal{S} = \{\boldsymbol{x}_j\}_{j=1}^n$ to maximize the empirical complete-task value defined below. A τ-approximate empirical maximizer, satisfying $\widehat{J}_n(\widehat{\boldsymbol{\theta}}) \ge \sup_{\boldsymbol{\theta}} \widehat{J}_n(\boldsymbol{\theta}) - \tau$, incurs an additional regret term $\tau \ge 0$. The sample budget counts prompts with conditional-value access rather than logged trajectories.

Joint value class. There are K agents with policy classes $\Pi_i = \{\pi_{\theta_i} : \theta_i \in \Theta_i\}$. For a prompt $x \sim \mathcal{D}$, the joint policy $\boldsymbol{\pi} = (\pi_{\theta_1}, \ldots, \pi_{\theta_K})$ produces a transcript Y and obtains a bounded reward $R(x, Y) \in [0, 1]$. Let

$$F_{\boldsymbol{\theta}}(x) \;:=\; \mathbb{E}[R(x, Y)] \in [0, 1], \qquad J(\boldsymbol{\theta}) \;:=\; \mathbb{E}_{x\sim\mathcal{D}}[F_{\boldsymbol{\theta}}(x)].$$

Given i.i.d. $x_1, \ldots, x_n \sim \mathcal{D}$ and access to these conditional expectations, define the empirical objective

$$\widehat{J}_n(\boldsymbol{\theta}) \;:=\; \frac{1}{n}\sum_{j=1}^{n} F_{\boldsymbol{\theta}}(x_j).$$

E.1 Per-agent sequence discrepancy and value Lipschitzness

The SARL argument applies to each agent conditional on its entering history.

Lemma E.1 (Sequence TV control for agent i)**.** *Let $T^{(i)}_{\max}$ denote the maximum sequence length for agent i. Under Assumption 3, let $L^{(i)}_{\mathrm{seq}} := T^{(i)}_{\max} L_{\mathrm{step}}$. Then, for any fixed prompt $\mathbf{x}$ and entering history $h = \mathbf{y}^{(<i)}$, the conditional full-segment laws satisfy*

$$\mathrm{TV}\big(\mathbb{P}_{\boldsymbol{\theta}_i}(\cdot \mid \mathbf{x}, h),\, \mathbb{P}_{\boldsymbol{\theta}'_i}(\cdot \mid \mathbf{x}, h)\big) \;\le\; L^{(i)}_{\mathrm{seq}} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}'_i\|_2.$$

Proof. Conditional on the entering history, the hybrid argument of Lemma D.1 applies to the token kernels of the segment. The policy Lipschitz assumption ensures that the resulting bound holds uniformly over $(\mathbf{x}, h)$. □

Lemma E.2 (Per-agent value Lipschitzness)**.** *For agent i, define*

$$f_{i,\boldsymbol{\theta}_i}(\mathbf{x}, h) = \mathbb{E}_{Y^{(i)}\sim\pi_{\boldsymbol{\theta}_i}(\cdot|\mathbf{x},h)} r_i(\mathbf{x}, Y^{(i)}, h).$$

Then for every $(\mathbf{x}, h)$,

$$|f_{i,\boldsymbol{\theta}_i}(\mathbf{x}, h) - f_{i,\boldsymbol{\theta}'_i}(\mathbf{x}, h)| \;\le\; L^{(i)}_{\mathrm{seq}} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}'_i\|_2.$$

Proof. The local reward lies in $[0, 1]$, so the difference of its expectations is bounded by the conditional total variation in Lemma E.1. The bound holds conditional on the entering history. Averaging over histories can still introduce dependence on upstream parameters. □

E.2 Per-agent covering numbers via parameter covers

Let $\mathcal{F}_i := \{f_{i,\theta_i} : \theta_i \in \Theta_i\}$ on the domain of allowed prompt–history pairs. Under the independent-policy assumption, histories are absent and the domain is $\mathcal{X}$. Denote covering numbers by $\mathcal{N}(\varepsilon, \cdot, \cdot)$ and $H(\varepsilon) := \log \mathcal{N}(\varepsilon, \cdot, \cdot)$.

Lemma E.3 (Agent Sample Complexity via Covering the Value Class)**.** *Let $\mathcal{N}(\varepsilon, \mathcal{F}_i, \|\cdot\|_\infty)$ denote the covering number of a function class $\mathcal{F}_i$ with respect to the ℓ_∞-norm, and let $H_{\mathcal{F}_i}(\varepsilon) := \log \mathcal{N}(\varepsilon, \mathcal{F}_i, \|\cdot\|_\infty)$. Then, under MARL setting and Assumption 2, the following holds:*

$$H_{\mathcal{F}_i}(\varepsilon) \le d_i \log\left(1 + \frac{2L^{(i)}_{\mathrm{seq}} B_i}{\varepsilon}\right).$$

Proof. Combine Lemma E.2 with the parameter-cover argument of Lemma D.3 and the Euclidean bound $(1 + 2B_i/\eta)^{d_i}$. When $L^{(i)}_{\mathrm{seq}} = 0$, one center suffices. □

Lemma E.4. *Let $\Theta \subset \mathbb{R}^d$ be a compact parameter set and let $\{f_\theta : \theta \in \Theta\}$ be a function class such that the parameterization is L-Lipschitz in the uniform (sup) norm, with $L > 0$:*

$$\|f_\theta - f_{\theta'}\|_\infty \;\le\; L\|\theta - \theta'\|_2, \quad \forall \theta, \theta' \in \Theta.$$

Define the k-ary aggregation map

$$\rho_k(z_1, \ldots, z_k) \;=\; \frac{1}{K}\sum_{i=1}^{k} z_i,$$

which is L_ρ-Lipschitz with respect to ℓ_1, with $L_\rho = 1/K$. Consider the aggregated function class

$$\mathcal{F}_k := \Big\{(z_1, \dots, z_k) \mapsto \rho_k(f_{\theta_1}(z_1), \dots, f_{\theta_k}(z_k)) : \theta_i \in \Theta\Big\}.$$

Then for $1 \le k \le K$ and every $\varepsilon > 0$,

$$\mathcal{N}(\varepsilon, \mathcal{F}_k, \|\cdot\|_\infty) \;\le\; \left[\mathcal{N}\left(\frac{K\varepsilon}{kL}, \Theta, \|\cdot\|_2\right)\right]^k.$$

Proof. By assumption, for each $\theta_i \in \Theta$ there exists u_i from an $(K\varepsilon/(kL))$-cover of Θ under $\|\cdot\|_2$ such that $\|f_{\theta_i} - f_{u_i}\|_\infty \le K\varepsilon/k$. For the tuples $\theta_{1:k}$ and $u_{1:k}$, the Lipschitz property of ρ_k with constant $L_\rho = 1/K$ gives

$$\Big\|\rho_k\big(f_{\theta_1}, \dots, f_{\theta_k}\big) - \rho_k\big(f_{u_1}, \dots, f_{u_k}\big)\Big\|_\infty \;\le\; L_\rho \sum_{i=1}^{k} \|f_{\theta_i} - f_{u_i}\|_\infty.$$

Since each difference is at most $K\varepsilon/k$, this yields

$$\Big\|\rho_k\big(f_{\theta_1}, \dots, f_{\theta_k}\big) - \rho_k\big(f_{u_1}, \dots, f_{u_k}\big)\Big\|_\infty \;\le\; \frac{1}{K} \cdot k \cdot \frac{K\varepsilon}{k} \;=\; \varepsilon.$$

Thus, the product cover of Θ^k induces an ε-cover of $\mathcal{F}_k$ in sup norm. Since a cover of Θ of size $\mathcal{N}(K\varepsilon/(kL), \Theta, \|\cdot\|_2)$ induces a cover of Θ^k of size $\big[\mathcal{N}(K\varepsilon/(kL), \Theta, \|\cdot\|_2)\big]^k$, the result follows. □

Corollary E.1 (Per-agent specialization)**.** *For agent i, let $\Theta_i \subset \mathbb{R}^{d_i}$ be contained in an ℓ_2-ball of radius B_i, and assume*

$$\|f_{i,\theta_i} - f_{i,\theta_i'}\|_\infty \;\le\; L_{\mathrm{seq},i}\, \|\theta_i - \theta_i'\|_2 \qquad \forall\, \theta_i, \theta_i' \in \Theta_i.$$

Then

$$H(\varepsilon,\, \mathcal{F}_i,\, \|\cdot\|_\infty) \;\le\; d_i \log\left(1 + \frac{2L_{\mathrm{seq},i} B_i}{\varepsilon}\right),$$

by Lemma E.3.

E.3 Proof of MARL Sample Complexity with Dependence

Proof. Fix $\varepsilon, \delta \in (0,1)$ and let $\mathcal{F}^{(K)} = \{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \prod_i \Theta_i\}$ be the class of complete conditional values. Although transcript generation is sequential, the learner maximizes the joint objective $\widehat{J}_n(\boldsymbol{\theta})$.

For any fixed prompt, replace the conditional segment kernels one at a time to form hybrid transcript laws. Two adjacent hybrids have the same entering-history distribution. Lemma E.1 bounds the change at segment i by $L_{\mathrm{seq},i}\|\theta_i - \theta_i'\|_2$, uniformly over that history. Appending the identical downstream kernels cannot increase total variation, even though those kernels condition on preceding outputs. The triangle inequality and bounded complete reward give

$$\|F_{\boldsymbol{\theta}} - F_{\boldsymbol{\theta}'}\|_\infty \le \sup_x \mathrm{TV}(\pi_{\boldsymbol{\theta}}(\cdot \mid x), \pi_{\boldsymbol{\theta}'}(\cdot \mid x)) \le \sum_{i=1}^{K} L_{\mathrm{seq},i} \|\theta_i - \theta_i'\|_2.$$

For $u > 0$, cover parameter block i at radius $u/(KL_{\mathrm{seq},i})$. A block with zero Lipschitz constant needs only one representative. The product of the covers has total value error at most u and therefore

$$H\Big(u, \mathcal{F}^{(K)}, \|\cdot\|_\infty\Big) \le \sum_{i=1}^{K} d_i \log\left(1 + \frac{2KL_{\mathrm{seq},i} B_i}{u}\right). \tag{E.3}$$

This argument controls the joint conditional value; it does not treat dependent local values as functions of their own parameters alone.

Apply the covering-number concentration argument from the SARL proof with cover radius $\varepsilon/8$ and net deviation $\varepsilon/4$. The complete class then has uniform deviation at most $\varepsilon/2$, except on an event of probability at most

$$2\exp\left\{\sum_{i=1}^{K} d_i \log\left(1+\frac{16KL_{\mathrm{seq},i}B_i}{\varepsilon}\right)-\frac{n\varepsilon^2}{8}\right\}.$$

Solving this failure-probability bound for n gives the sufficient budget

$$U_{\mathrm{D}}(\varepsilon,\delta) := \frac{8}{\varepsilon^2}\left[\sum_{i=1}^{K} d_i \log\left(1+\frac{16KL_{\mathrm{seq},i}B_i}{\varepsilon}\right)+\log\frac{2}{\delta}\right], \tag{E.4}$$

which ensures, for every integer $n \geq \lceil U_{\mathrm{D}} \rceil$, that with probability at least $1-\delta$,

$$\sup_{\boldsymbol{\theta}} |J(\boldsymbol{\theta}) - \widehat{J}_n(\boldsymbol{\theta})| \leq \varepsilon/2. \tag{E.5}$$

On this event, the complete-return empirical maximizer with tolerance τ satisfies

$$\sup_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) - J(\widehat{\boldsymbol{\theta}}) \leq \sup_{\boldsymbol{\theta}} \widehat{J}_n(\boldsymbol{\theta}) - \widehat{J}_n(\widehat{\boldsymbol{\theta}}) + \varepsilon \leq \varepsilon + \tau. \tag{E.6}$$

For exact ERM, take $\tau = 0$.

This proves Theorem 3.2 directly. The chosen per-segment resolution u/K controls the joint cover error and is already included in (E.3). The concentration event concerns a single class of complete returns, each in $[0,1]$; it does not require K separate statistical errors of size ε/K. Hence the same ε^{-2} concentration scale as SARL suffices. This argument includes dependence of downstream outputs on upstream segments, but does not assert that greedy local training attains the joint empirical optimum. □

E.4 Proof of Sample Complexity for MARL under Independence

The proof of Theorem 3.3 uses policy and reward factorization.

Proof. For each agent i, define the conditional-value class

$$\mathcal{F}_i := \left\{ f_{i,\theta_i}(x) := \mathbb{E}_{Y^{(i)}\sim\pi_{\theta_i}(\cdot|x)} r_i(x, Y^{(i)}) : \theta_i \in \Theta_i \right\} \subset [0,1]^{\mathcal{X}}.$$

The assumed policy factorization, local rewards, and Cartesian parameter set give the complete value class

$$\mathcal{G} := \left\{ g_{\theta_{1:K}}(x) = \frac{1}{K}\sum_{i=1}^{K} f_{i,\theta_i}(x) : \theta_i \in \Theta_i \ \forall i \right\}.$$

Let $J(g) = \mathbb{E}g(X)$ and $\widehat{J}_n(g) = n^{-1}\sum_{j=1}^{n} g(X_j)$. If

$$\sup_{g\in\mathcal{G}} |J(g) - \widehat{J}_n(g)| \leq \eta, \tag{E.7}$$

then an empirical maximizer $\widehat{g}$ with tolerance τ satisfies

$$\sup_{g\in\mathcal{G}} J(g) - J(\widehat{g}) \leq \sup_{g\in\mathcal{G}} \widehat{J}_n(g) - \widehat{J}_n(\widehat{g}) + 2\eta \leq \tau + 2\eta.$$

It therefore suffices to establish (E.7) with $\eta = \varepsilon/2$.

For any parameter tuple, apply the triangle inequality and then take the supremum to obtain

$$\sup_{g\in\mathcal{G}} |J(g) - \widehat{J}_n(g)| \leq \frac{1}{K}\sum_{i=1}^{K} \sup_{\theta_i\in\Theta_i} |J(f_{i,\theta_i}) - \widehat{J}_n(f_{i,\theta_i})|. \tag{E.8}$$

The SARL covering-number proof, applied to local class i with failure probability δ/K, shows that

$$n \geq \frac{8}{\varepsilon^2}\left[d_i \log\left(1+\frac{16L_{\mathrm{seq},i}B_i}{\varepsilon}\right)+\log\frac{2K}{\delta}\right] \tag{E.9}$$

is sufficient for its uniform deviation to be at most $\varepsilon/2$. A union bound over the K local events yields, with probability at least $1-\delta$,

$$\sup_{\theta_i \in \Theta_i} |J(f_{i,\theta_i}) - \widehat{J}_n(f_{i,\theta_i})| \le \varepsilon/2 \qquad \text{for every } i \in [K]. \tag{E.10}$$

The events may be correlated through the shared prompts; the union bound does not require independence between them.

Set $\widetilde{d} = \max_i d_i$ and $\gamma = \max_i L_{\mathrm{seq},i} B_i$. The budget

$$U_{\mathrm{I}}(\varepsilon, \delta) = \frac{8}{\varepsilon^2}\left[\widetilde{d}\log\left(1 + \frac{16\gamma}{\varepsilon}\right) + \log\frac{2K}{\delta}\right]$$

is at least the right-hand side of (E.9) for each i. Thus $n \ge \lceil U_{\mathrm{I}} \rceil$ ensures the simultaneous event. Combining (E.8) and (E.10) gives the required global deviation and regret at most $\varepsilon + \tau$. Exact maximization of each local empirical objective also maximizes their average because the parameter set is Cartesian. The $\log K$ confidence term arises from controlling all agents simultaneously; this proof does not establish that it is necessary. □

F Proof of Relative Efficiency

This section compares the sufficient sample budgets for SARL and MARL. Their ratios describe the guarantees established here, not the unknown minimal sample requirements.

F.1 Proof of Proposition 3.4

Proof. Theorems 3.1 and 3.3 give the sufficient budgets

$$U_{\mathrm{S}}(\varepsilon, \delta) = \frac{8}{\varepsilon^2}(A + L), \tag{F.1}$$

$$U_{\mathrm{I}}(\varepsilon, \delta) = \frac{8}{\varepsilon^2}(M + L_K), \tag{F.2}$$

where

$$A := d\log\left(1 + \frac{16 L_{\mathrm{seq}} B}{\varepsilon}\right), \qquad L := \log(2/\delta),$$

$$M := \widetilde{d}\log\left(1 + \frac{16\gamma}{\varepsilon}\right), \qquad L_K := \log(2K/\delta).$$

Here $L_{\mathrm{seq}} = T_{\max} L_{\mathrm{step}}$, $\widetilde{d} = \max_i d_i$, and $\gamma = \max_i (L_{\mathrm{seq},i} B_i)$. For each theorem, an integer sample count at least the ceiling of the corresponding budget suffices. Since $L_K = L + \log K$, the exact ratio of these real-valued budgets is

$$\frac{U_{\mathrm{I}}(\varepsilon, \delta)}{U_{\mathrm{S}}(\varepsilon, \delta)} = \frac{M + L_K}{A + L}.$$

In particular, $M + \log K \le A$ is sufficient for $U_{\mathrm{I}} \le U_{\mathrm{S}}$ and hence for the same ordering of their ceilings. This is a comparison of sufficient budgets; dividing two upper bounds on minimal sample complexities would not establish such an ordering.

If the capacity bounds $\widetilde{d} \le d$ and $\gamma \le L_{\mathrm{seq}} B$ hold, monotonicity of $\log(1+x)$ gives $M \le A$. These assumptions alone do not remove the extra $\log K$ in the simultaneous confidence term, so they do not imply an unconditional advantage.

For the homogeneous case, assume $d_i = d/K$, $T_{\max,i} = T_{\max}/K$, and $B_i \le B$ for all i, with the same L_{step}. Then $\widetilde{d} = d/K$ and $\gamma \le L_{\mathrm{seq}} B/K$. Therefore,

$$M \le \frac{d}{K}\log\left(1 + \frac{16 L_{\mathrm{seq}} B}{K\varepsilon}\right) \le \frac{A}{K}.$$

Substitution into the exact budget ratio yields

$$\begin{aligned}\frac{U_{\mathrm{I}}(\varepsilon,\delta)}{U_{\mathrm{S}}(\varepsilon,\delta)} &\le \frac{A/K + L + \log K}{A + L} \\ &= \frac{1}{K} + \frac{(1-1/K)\log(2/\delta) + \log K}{d\log(1 + 16L_{\mathrm{seq}}B/\varepsilon) + \log(2/\delta)}.\end{aligned} \tag{F.3}$$

For fixed K and δ, the confidence correction tends to zero as $A \to \infty$; an asymptotic statement with a correction $o(1/K)$ instead requires $K[(1-1/K)L + \log K]/(A+L) \to 0$. No unconditional $1/K$ ratio of sample complexities follows from these certificates. □

F.2 Proof of Proposition 3.5

Proof. Theorems 3.1 and 3.2 establish the sufficient budgets

$$U_{\mathrm{S}}(\varepsilon,\delta) = \frac{8}{\varepsilon^2}(A + L), \tag{F.4}$$

$$U_{\mathrm{D}}(\varepsilon,\delta) = \frac{8}{\varepsilon^2}(S_K + L), \tag{F.5}$$

where

$$A := d\log\left(1 + \frac{16L_{\mathrm{seq}}B}{\varepsilon}\right), \qquad S_K := \sum_{i=1}^{K} d_i \log\left(1 + \frac{16KL_{\mathrm{seq},i}B_i}{\varepsilon}\right), \qquad L := \log(2/\delta).$$

Both concentration arguments use one joint value class and the same failure probability δ. Dividing the two budgets gives

$$\frac{U_{\mathrm{D}}(\varepsilon,\delta)}{U_{\mathrm{S}}(\varepsilon,\delta)} = \frac{S_K + L}{A + L}.$$

Because $L > 0$, this ratio is well defined even when either entropy term vanishes, and $U_{\mathrm{D}} \le U_{\mathrm{S}}$ is equivalent to $S_K \le A$. The same ordering holds for the integer ceilings, although two distinct real-valued budgets can have equal ceilings.

If $\sum_i d_i = d$, $L_{\mathrm{seq},i} = L_{\mathrm{seq}}/K$, and $B_i \le B$, then

$$S_K = \sum_i d_i \log\left(1 + \frac{16L_{\mathrm{seq}}B_i}{\varepsilon}\right) \le \sum_i d_i \log\left(1 + \frac{16L_{\mathrm{seq}}B}{\varepsilon}\right) = A.$$

Equality holds if every $B_i = B$. In particular, for $K = 1$ and matched parameters, the two budgets coincide. The comparison concerns sufficient prompt budgets for regret relative to each class optimum. Dependence therefore need not enlarge this certificate under the stated matching conditions. Differences in attainable performance and the behavior of particular optimizers remain separate questions. □

G Imperfect Task Alignment

Reward discrepancy and differences between policy-class optima both affect the sufficient sample budgets.

G.1 Proof of Theorem 3.6

Proof. Fix the modular policy class Π_{M} used both to optimize the surrogate and to evaluate its target-reward regret. Let $R, \overline{R}$ take values in $[0,1]$, and write $J_R(\pi) := \mathbb{E}[R(x,y)]$ and $J_{\overline{R}}(\pi) := \mathbb{E}[\overline{R}(x,y)]$ for the same prompt distribution and policy-induced transcript distribution. The uniform reward discrepancy $\sup_{x,y} |R(x,y) - \overline{R}(x,y)| \le \alpha$ implies

$$\begin{aligned}&\forall \pi \in \Pi_{\mathrm{M}}: \quad |J_R(\pi) - J_{\overline{R}}(\pi)| \le \alpha, \\ &\left|\sup_{\pi\in\Pi_{\mathrm{M}}} J_R(\pi) - \sup_{\pi\in\Pi_{\mathrm{M}}} J_{\overline{R}}(\pi)\right| \le \alpha.\end{aligned} \tag{G.1}$$

Consequently, for any $\widehat{\pi} \in \Pi_{\mathrm{M}}$,

$$\sup_{\pi \in \Pi_{\mathrm{M}}} J_R(\pi) - J_R(\widehat{\pi}) \leq \left[\sup_{\pi \in \Pi_{\mathrm{M}}} J_{\overline{R}}(\pi) - J_{\overline{R}}(\widehat{\pi}) \right] + 2\alpha. \tag{G.2}$$

Thus a surrogate excess reward at most

$$\beta := \varepsilon - 2\alpha > 0 \tag{G.3}$$

suffices for target excess reward at most ε in this same class.

To invoke the independent-task certificate, assume the exact conditional surrogate value factorizes on the product parameter set:

$$\mathbb{E}_{y \sim \pi_{\theta_{1:K}}(\cdot|x)}[\overline{R}(x,y)] = \frac{1}{K} \sum_{i=1}^{K} f_i(x; \theta_i), \qquad \theta_{1:K} \in \prod_{i=1}^{K} \Theta_i,$$

where $f_i \in [0,1]$ satisfies the boundedness and parameter-Lipschitz conditions of Theorem 3.3. This is a condition on policy values, not merely on the algebraic additivity of $\overline{R}$. For the common i.i.d. prompt sample $x_1, \ldots, x_n$, set $J_i(\theta_i) = \mathbb{E}_x f_i(x; \theta_i)$ and $\widehat{J}_{i,n}(\theta_i) = n^{-1} \sum_{j=1}^{n} f_i(x_j; \theta_i)$. Then

$$J_{\overline{R}}(\theta_{1:K}) = \frac{1}{K} \sum_i J_i(\theta_i), \qquad \widehat{J}_{\overline{R},n}(\theta_{1:K}) = \frac{1}{K} \sum_i \widehat{J}_{i,n}(\theta_i).$$

The triangle inequality gives

$$\begin{aligned} \sup_{\theta_{1:K}} |J_{\overline{R}}(\theta_{1:K}) - \widehat{J}_{\overline{R},n}(\theta_{1:K})| &\leq \frac{1}{K} \sum_i \sup_{\theta_i} |J_i(\theta_i) - \widehat{J}_{i,n}(\theta_i)| \\ &\leq \max_i \sup_{\theta_i} |J_i(\theta_i) - \widehat{J}_{i,n}(\theta_i)|. \end{aligned} \tag{G.4}$$

A union bound controls all agents' empirical deviations simultaneously, even though they are computed from shared prompts. The proof of Theorem 3.3, with accuracy β, ensures that the right-hand side is at most $\beta/2$ with probability at least $1 - \delta$ whenever

$$\begin{aligned} n &\geq \lceil U_{\mathrm{I}}(\beta, \delta) \rceil, \\ U_{\mathrm{I}}(\beta, \delta) &= \frac{8}{\beta^2} \left[\widetilde{d} \log\left(1 + \frac{16\gamma}{\beta}\right) + \log\left(\frac{2K}{\delta}\right) \right], \end{aligned}$$

where $\widetilde{d} = \max_i d_i$ and $\gamma = \max_i (L_{\mathrm{seq},i} B_i)$. An exact empirical maximizer $\widehat{\theta}_{1:K}$ can be obtained by maximizing each $\widehat{J}_{i,n}$ over Θ_i, because the parameter set is a product and the objective factorizes. On this event, the empirical maximization inequality gives

$$\sup_{\theta_{1:K}} J_{\overline{R}}(\theta_{1:K}) - J_{\overline{R}}(\widehat{\theta}_{1:K}) \leq 2 \sup_{\theta_{1:K}} |J_{\overline{R}}(\theta_{1:K}) - \widehat{J}_{\overline{R},n}(\theta_{1:K})| \leq \beta.$$

Combining this with (G.2) proves the sufficient bound $N_{\mathrm{MARL}}(\varepsilon, \delta) \leq \lceil U_{\mathrm{I}}(\varepsilon - 2\alpha, \delta) \rceil$ for the fixed-class target regret.

For comparison to a different SARL class Π_{S}, define its possible target-value advantage by

$$a := \max\left\{ 0, \sup_{\pi \in \Pi_{\mathrm{S}}} J_R(\pi) - \sup_{\pi \in \Pi_{\mathrm{M}}} J_R(\pi) \right\}.$$

For every $\widehat{\pi} \in \Pi_{\mathrm{M}}$, adding this comparator gap to (G.2) bounds its regret relative to Π_{S} by $a + 2\alpha$ plus its surrogate regret within Π_{M}. Consequently, when $\varepsilon - a - 2\alpha > 0$, the same proof with $\beta = \varepsilon - a - 2\alpha$ gives the sufficient budget $\lceil U_{\mathrm{I}}(\varepsilon - a - 2\alpha, \delta) \rceil$. The reward discrepancy alone does not control a. □

G.2 Proof of Proposition 3.7

Proof. Let $\beta = \varepsilon - a - 2\alpha > 0$, where a is the comparator gap defined above (and $a = 0$ for a common comparator class). The sufficient budgets from Theorems 3.1 and 3.6 are

$$U_{\mathrm{S}}(\varepsilon, \delta) = \frac{8(A + L)}{\varepsilon^2}, \qquad U_{\mathrm{I}}(\beta, \delta) = \frac{8(C + L_K)}{\beta^2},$$

where

$$A := d \log\left(1 + \frac{16 L_{\mathrm{seq}} B}{\varepsilon}\right), \qquad C := \widetilde{d} \log\left(1 + \frac{16\gamma}{\beta}\right),$$
$$L := \log(2/\delta), \qquad L_K := \log(2K/\delta).$$

Dividing these budgets gives

$$\frac{U_{\mathrm{I}}(\beta, \delta)}{U_{\mathrm{S}}(\varepsilon, \delta)} = \frac{\varepsilon^2}{\beta^2} \frac{C + L_K}{A + L}.$$

For $A, C > 0$, this is also

$$\left(\frac{\varepsilon}{\beta}\right)^2 \underbrace{\frac{C}{A}}_{\kappa_d \kappa_\ell} \underbrace{\frac{1 + L_K/C}{1 + L/A}}_{\mathcal{C}_\alpha(\varepsilon, \delta)},$$

where

$$\kappa_d = \frac{\widetilde{d}}{d}, \quad \kappa_\ell = \frac{\log(1 + 16\gamma/\beta)}{\log(1 + 16 L_{\mathrm{seq}} B/\varepsilon)}.$$

It follows that the condition

$$C + L_K \le \left(\frac{\beta}{\varepsilon}\right)^2 (A + L)$$

is sufficient for $U_{\mathrm{I}}(\beta, \delta) \le U_{\mathrm{S}}(\varepsilon, \delta)$ and the corresponding ordering of their ceilings. Equivalently, when $A, C > 0$, this condition can be written as

$$\kappa_d \kappa_\ell \mathcal{C}_\alpha(\varepsilon, \delta) \le (\beta/\varepsilon)^2.$$

When $a = 0$, the right-hand side is $(1 - 2\alpha/\varepsilon)^2$. The unfactored inequality also applies if either capacity term vanishes. The confidence correction remains part of the sufficient condition. The result compares the guarantees established here; it gives neither a necessary condition for a sample-efficiency advantage nor an ordering of the unknown minimal sample complexities. □

H Experiment Details

H.1 Testbed

The experiments were conducted on an AMD-based system running Windows 11 (build 26200), with 96 physical CPU cores, 192 logical processors, and 127.69 GiB of OS-visible RAM. All GPU experiments used a single NVIDIA GeForce RTX 5090 D v2, operating in WDDM mode with NVIDIA driver 595.79. The GPU exposed 23.88 GiB of addressable memory, with an experimental allocation cap of 17.91 GiB (75%). The software environment comprised Python 3.12.14, PyTorch 2.10.0+cu128, the CUDA 12.8 runtime, and NVRTC 12.8

H.2 Synthetic Noisy Arithmetic Tasks

The regression task maps local inputs $\mathbf{x}_i$ to scalar outputs:

$$y_i = \mathbf{w}_i^\top \mathbf{x}_i + \lambda \cdot \mathrm{average}\big(y^{(<i)}\big) + \xi_i, \qquad \xi_i \sim \mathcal{N}(0, 1). \tag{H.1}$$

The empty-prefix term is zero. The endpoints $\lambda = 0$ and $\lambda = 1$ give the independent and dependent tasks, respectively. Dependence permits earlier predictions to affect later outputs but does not prescribe greedy training. The coupling $\lambda \in [0, 1]$ is distinct from reward discrepancy α.

H.3 Hyperparameter Tuning

The reported experiments used Adam, with learning rates selected by grid search on an independent validation set. Test performance was excluded from selection; the search grid and selected rates are unspecified.

H.4 Coupling, Agent Count, and GSM8K Tasks

Synthetic comparisons. Figure 1(c) was designed to compare $\lambda = 0.1$ (Strong) with $\lambda = 1$ (Weak), describing the intended compatibility of local decomposition with weaker or stronger interactions. The agent-count sweep uses $K \in \{1, 2, 3, 4\}$ under both task structures, with total model capacity reported as fixed by adjusting per-agent dimensionality.

GSM8K constructions. Both task constructions use `Qwen2.5-1.5B-Instruct` and Adam. Training uses GRPO with group size eight, as detailed in Section H.6. Independent tasks concatenate two unrelated questions, one per MARL agent. Dependent tasks assign reasoning to Agent 1 and final-answer generation from that reasoning to Agent 2. SARL trains one model on the complete task; MARL trains two Qwen-based agents on their assigned segments. Neither construction measures reward discrepancy α.

H.5 Proposed Synthetic Protocol

Data and sample units. The proposed supervised regression study uses $D = 48$ total features, $K \in \{1, 2, 3, 4\}$ outputs, and $p = D/K$ local features. Five trials use seeds $s \in \{0, 1, 2, 3, 4\}$. Independent inputs $\mathbf{x}_i \sim \mathcal{N}(0, I_p)$ and innovations $\xi_i \sim \mathcal{N}(0, 1)$ generate

$$y_i = \mathbf{w}_i^\top \mathbf{x}_i + \frac{\lambda}{i-1} \sum_{j<i} y_j + \xi_i \quad (i > 1), \qquad y_1 = \mathbf{w}_1^\top \mathbf{x}_1 + \xi_1.$$

Task vectors drawn from $\mathcal{N}(0, I_p)$ are normalized to unit norm and fixed across sample sizes and learners. Independent random streams control weights, training, validation, test data, and initialization. Learners share data within a trial; coupling sweeps also share weights, features, and innovations across λ. Trials are independent.

Nested training sizes are $n \in \{50, 100, 200, 400, 800, 1600, 3200\}$, with 2,048 validation and 16,384 test pairs. One sample is a complete input–trajectory pair containing all K labels. A separate global comparator is fitted at each K, since changing K changes the output task and local dimension.

Architecture and capacity. MLPs have one hidden ReLU layer, linear outputs, biases, Xavier-uniform weights, zero initial biases, and no dropout. The global learner maps $\mathbb{R}^D$ to $\mathbb{R}^K$; role i predicts one scalar from $q_i = p$ inputs for independent policies or $q_i = p + i - 1$ with preceding outputs. For widths h_{S} and h_i, parameter counts are

$$P_{\mathrm{S}} = h_{\mathrm{S}}(D + K + 1) + K, \qquad P_{\mathrm{M}} = \sum_{i=1}^{K} \{h_i(q_i + 2) + 1\}.$$

With $P^\star = 32{,}768$, choose $h_{\mathrm{S}} = \lfloor (P^\star - K)/(D + K + 1) \rfloor$ and $h_i = \lfloor (P^\star - K)/\sum_j (q_j + 2) \rfloor$. Integer widths approximately match parameter counts, not function classes. The independent comparison uses $\lambda = 0$ and history-independent roles. A separate sweep over $\lambda \in \{0, 0.1, 0.5, 1\}$ fixes the history-conditioned architecture; even at $\lambda = 0$, this does not satisfy the independent-policy premise of Theorem 3.3.

Training and selection. The global learner minimizes complete-output MSE. The stagewise learner fits local squared error on reference prefixes, freezing roles in order; deployment uses generated prefixes. A third learner jointly minimizes complete generated-output MSE with the same modular architecture and gradients through preceding predictions. Both modular learners share initialization within each trial.

Adam uses $(\beta_1, \beta_2) = (0.9, 0.999)$, stabilizer 10^{-8}, zero weight decay, batch size $\min(128, n)$, gradient-norm clipping at 1, epoch-wise shuffling, and no schedule. The search crosses learning rates

$\{10^{-4}, 3 \cdot 10^{-4}, 10^{-3}, 3 \cdot 10^{-3}\}$ with checkpoints at $\{20, 50, 100, 200\}$ epochs. Roles share each candidate rate and epoch count. Selection minimizes complete validation MSE under deployment histories; ties favor fewer epochs, then a smaller rate. Test data are used only for final evaluation. Update counts and compute are recorded separately because equal data passes need not imply equal cost.

Metrics and diagnostics. To valuate, we set aside an independent test set and evaluate the prediction with the standard MSE metric.

H.6 GSM8K Protocol

We evaluate the single-agent and decomposed constructions on GSM8K using Group Relative Policy Optimization (GRPO). All compared policies use the same GRPO objective, with eight sampled completions per reward group ($N_{\text{grp}} = 8$). The experiment provides a practical RL comparison; it is not a direct instantiation of the conditional-value PAC analysis or the reference-demonstration analysis in Section 3.4, and our theory does not establish GRPO convergence or rollout sample complexity.

Data and adaptation. We use `Qwen2.5-1.5B-Instruct` with its tokenizer and chat template on the official GSM8K train/test splits. Dataset and model versions are fixed across learners.[1] Each of five seeds $s \in \{0, 1, 2, 3, 4\}$ shuffles the training questions, with 512 questions reserved for validation. The official test split is used only for final evaluation.

For the independent construction, questions within each split are randomly permuted and paired consecutively without reuse; an unmatched final question is omitted. Each complete example therefore contains two independent GSM8K questions. The dependent construction contains one question per complete example and decomposes its solution into reasoning and final-answer generation. Nested training pools contain $n \in \{50, 100, 200, 400, 800, 1600, 5000\}$ complete examples where available. All learners in a trial use the same split and question pairing.

Here, n denotes the number of distinct complete training examples available to the learner, rather than the number of sampled trajectories. For every GRPO training prompt, eight responses are sampled to form one reward group. Training examples, generated rollouts, and reward evaluations are therefore distinct counts. Independent complete examples also contain two underlying GSM8K questions each.

The backbone is frozen and adaptation is restricted to the query and value projection matrices in each transformer block through rank-r updates

$$W \mapsto W + 2BA,$$

with adapter-input dropout 0.05, Kaiming-uniform initialization of A, and zero initialization of B. SARL uses one rank-16 adapter, whereas MARL uses two distinct rank-8 adapters. No biases or other parameters are trainable. For the same target matrices, a rank-r update has $r(d_{\text{in}} + d_{\text{out}})$ trainable parameters, giving the two configurations equal total trainable parameter counts. This matching does not equate their policy classes, memory use, rollout cost, or inference cost.

Task constructions and prompts. Reference GSM8K solutions are used to obtain the ground-truth numerical answers for reward computation and evaluation; they are not used as token-level supervised training targets. Calculator annotations of the form «...» are removed when processing the reference answers. All configurations use the system message *Solve the given math task carefully* and the following user prompts:

> **Single question / independent role:**
> Question: [question]. Give your reasoning, then write the final answer as `#### [number]`.
> **Independent SARL:**
> Question 1: [question 1]. Question 2: [question 2]. Solve both in order. Use headings `Solution 1:` and `Solution 2:`. Under each heading give reasoning and end with `#### [number]`.

[1] Primary resources: the official GSM8K repository and the Qwen model card.

Dependent thinker:
Question: [question]. Give reasoning to support a final solution. Do not add a #### final-answer line.

Dependent solver:
Question: [question]. Reasoning: [reasoning]. Give only the final answer as #### [number].

For independent tasks, SARL solves the two questions jointly, whereas the two MARL roles solve one question each. For dependent tasks, SARL generates a complete reasoning-and-answer trajectory. MARL first generates a reasoning trace with the thinker and then conditions the solver on that generated trace to produce the final answer. Thus the solver is evaluated on generated rather than reference reasoning.

GRPO optimization and rewards. All policies are trained using the standard clipped GRPO objective. For each training prompt x, the behavior policy samples a group $\{y_g\}_{g=1}^{N_{\text{grp}}}$ of $N_{\text{grp}} = 8$ completions. Let R_g denote the reward of completion g. The group-relative advantage is obtained by centering and normalizing rewards within the group,

$$\widehat{A}_g = \frac{R_g - \text{mean}(R_{1:N_{\text{grp}}})}{\text{std}(R_{1:N_{\text{grp}}}) + \epsilon_{\text{adv}}},$$

where $\epsilon_{\text{adv}} > 0$ stabilizes the denominator. Policy updates use the clipped importance-ratio objective, together with the implementation's KL regularization toward the reference policy. The same GRPO configuration and group size are used across the compared SARL and MARL configurations; total rollout accounting depends on how role completions and downstream continuations are grouped.

Rewards are based on GSM8K numerical-answer correctness. For a completed single-question solution, the reward is one when its parsed final numerical answer equals the reference answer and zero otherwise. For the independent SARL construction, the complete-example reward is the average correctness of the two questions,

$$R_{\text{indep}} = \frac{1}{2}\left(\mathbf{1}\{\widehat{a}_1 = a_1^\star\} + \mathbf{1}\{\widehat{a}_2 = a_2^\star\}\right),$$

while each independent MARL role receives the correctness reward for its assigned question.

For the dependent construction, the primary complete-task reward is the correctness of the final numerical answer produced from the generated reasoning history. This preserves the downstream dependence between the thinker and solver: errors in the generated reasoning can affect the reward through the solver's final prediction. Role-specific credit assignment and any auxiliary reward must be distinguished from this complete-task reward because they determine how each role is trained.

Rollouts, optimization, and resource accounting. Each GRPO group contains eight sampled completions. This group size alone does not determine the number of complete multi-role trajectories. Sampling is used during GRPO training; evaluation uses deterministic decoding. The same GRPO clipping, KL-regularization, optimizer, learning-rate selection procedure, and stopping criterion are used for the corresponding SARL and MARL comparisons unless otherwise stated.

Because a fixed number of distinct training prompts does not imply a fixed number of generated tokens or identical computational cost, unique training examples, rollout trajectories, reward evaluations, generated tokens, optimizer updates, and elapsed GPU time must be distinguished. In particular, the training-pool size n is not rollout sample complexity.

The context limit is 4,096 tokens, including prompts and generated responses. Independent tasks allow 2,048 generated tokens for SARL and 1,024 tokens per MARL role. Dependent tasks allow 1,152 generated tokens for SARL and 1,024 and 128 tokens for the thinker and solver, respectively. These caps match the maximum output-token allowance but do not imply equal realized computation.

Evaluation and history diagnostics. The answer parser extracts the numerical string after the last #### marker in each required answer block, removes whitespace, thousands separators, and a leading currency symbol, and compares signed integers and finite decimals after numerical normalization. Missing markers, missing answer blocks, and nonnumeric outputs are counted as incorrect.

For independent tasks, Figure 1(e) reports both-correct accuracy: an example is correct only when both answers are correct. Mean per-question accuracy is a separate diagnostic. For dependent tasks, the primary metric is final-answer accuracy when the solver conditions on generated reasoning. As a history-sensitivity diagnostic, we additionally evaluate the same solver after replacing generated reasoning with the reference reasoning. This intervention is excluded from the deployment score and is not identified with the policy-TV discrepancy in Theorem 3.8. Likewise, terminal-answer correctness does not directly instantiate the normalized prefix-reward decomposition in Theorem 3.8 without an additional reward decomposition and scaling. Neither GSM8K construction directly measures the worst-case reward discrepancy α.

I Additional Discussion

I.1 Effect of Parameter Sharing on the Dimension Terms

Agents may share an encoder, transformer layers, or embedding tables. The dimensions $\{d_i\}_{i=1}^K$ count the free trainable coordinates in each agent's parameter space Θ_i; frozen parameters are excluded. These dimensions, together with the radius and Lipschitz bounds, control the covering numbers; they are not themselves covering numbers.

Shared vs. non-shared parameters. A shared frozen backbone is compatible with separate trainable agent parameters. Tying *trainable* parameters instead restricts the joint class to a subset of the product class. The joint covering argument can still give a conservative bound for this subset, but a reduced dimension requires an explicit shared-coordinate parameterization and corresponding Lipschitz bounds. The independent-subtask theorem assumes a Cartesian product of trainable parameter sets, so that separately chosen agent policies form a feasible joint policy; it does not automatically apply to tied trainable blocks.

Impact on sample complexity. The independent-subtask budget U_{I} in Theorem 3.3 uses $\widetilde{d} = \max_i d_i$ under the specified policy and parameter factorization. A shared frozen backbone does not add trainable coordinates. Trainable parameter tying can restrict capacity, but also changes attainable rewards and can prevent separate empirical maximization. Thus it need not improve the relative guarantee. The dependent-subtask joint bound U_{D} remains a sufficient budget for a restricted class when empirical maximization is over that class; its dimension term is not automatically reduced by replacing a sum with a union of parameter subsets.

Effect on structural conclusions. Our comparison of sufficient budgets depends on subtask independence, sequential dependence, and task alignment within the stated policy classes. Parameter sharing must preserve the factorization assumptions to retain the independent-subtask result. Sharing can affect both dimension terms and expressivity, so it can also change the class gap when comparing to SARL. The joint budget U_{D} contains the per-segment cover resolution inside logarithms, without an external K^2 factor.

Heterogeneous agents. Agent dimensions may differ, including when agents share a frozen backbone. Tied trainable parameters require the joint-class interpretation above rather than independent local optimization.

I.2 Practical Estimation of the Alignment Factor α

The alignment factor α (Eq. (3.1)) measures the worst-case discrepancy between the original unified reward $R(x, y)$ and the reward induced by the MARL-style independent decomposition:

$$\alpha \;=\; \sup_{x,y} \left| R(x,y) \;-\; \frac{1}{K} \sum_{i=1}^{K} r_i\Big(x, y^{(i)}\Big) \right|. \tag{I.1}$$

A small value of α indicates uniformly small reward distortion, whereas a large α weakens the transfer guarantee and can increase its sufficient budget (Theorem 3.6 and Proposition 3.7). Two practical diagnostics can detect large discrepancies in LLM-based RL settings, but neither certifies an upper bound on α.

Monte Carlo estimation. Given a collection of rollouts $\mathcal{S} = \{(x_j, y_j)\}_{j=1}^n$ generated from one or more policies, an observed lower bound on α can be obtained via

$$\widehat{\alpha}_{\mathrm{MC}} \;=\; \max_{(x_j,y_j)\in\mathcal{S}} \left| R(x_j,y_j) \;-\; \frac{1}{K} \sum_{i=1}^{K} r_i\Big(x_j, y_j^{(i)}\Big) \right|. \tag{I.2}$$

With exact reward evaluations, $\widehat{\alpha}_{\mathrm{MC}} \le \alpha$; unobserved prompt–output pairs can have larger discrepancies. A high quantile (e.g., the 95th percentile) describes typical discrepancies under the sampling distribution, but also does not upper-bound the supremum. These diagnostics can be updated online as additional rollouts are collected.

Gradient-based maximization. A complementary approach seeks to approximate the worst-case discrepancy

$$\Delta(x, y) \;=\; \left| R(x, y) \;-\; \frac{1}{K} \sum_{i=1}^{K} r_i\Big(x, y^{(i)}\Big) \right|, \tag{I.3}$$

by performing gradient ascent on (x, y) (or on y conditioned on x) with respect to $\Delta(x, y)$ when a differentiable relaxation is available. This method can reveal adversarial cases or rare task instances with large discrepancies. Rewards evaluated on valid discrete candidates again give lower bounds on α; finding a local maximum does not certify the global supremum.

Interpreting the diagnostics. Both procedures can help diagnose poor reward decompositions. Neither certifies the small global discrepancy required by the alignment theorem. Using the theorem's numerical budget requires a valid upper bound on α, established separately from these observed maxima or quantiles. Moreover, small reward discrepancy does not remove a gap between the modular and SARL policy classes.

I.3 Covering-Number Bounds

Covering-number PAC bounds can be loose in absolute magnitude. Their role here is to expose how structural quantities enter the SARL and MARL guarantees.

Interpreting sufficient budgets. The closed-form *sufficient budgets* U_{S}, U_{D}, and U_{I} expose dependence on agent dimensionality d_i, the number of agents K, and decomposition structure (independent vs. dependent); the reward-transfer analysis accounts for alignment level α. These bounds are not numerically tight estimates of the samples required to train LLM-based RL systems. Ratios of these budgets are not ratios of minimum sample complexities, and loose upper bounds alone cannot establish an intrinsic architectural advantage or penalty. Theorem 3.2 uses the direct joint-cover budget; a finer per-segment cover does not require a separate concentration event at accuracy ε/K for each segment.

Relation to alternative complexity measures. Alternative measures, including Rademacher complexity, can also support generalization analysis. Uniform covering numbers make the dimension, radius, and Lipschitz dependence explicit without specifying a prompt distribution. This choice does not rule out using other measures to capture decomposition or alignment. The reward-transfer argument involving α is separate from the choice of concentration or complexity bound.

I.4 Theoretical Quantities and LLM–RL Training

The bounds depend on the dimensions d_i, d, radius parameters B_i, B, Lipschitz constants $L_{\text{step}}, L_{\text{seq}}$, and sequence length T. These quantities relate to LLM–RL training choices as follows.

Effective dimensions d_i and d. In our analysis, d_i denotes the number of free trainable coordinates of agent i's policy class, not the total number of parameters in a full LLM. In practice, this corresponds to the number of free parameters in the *trainable subset* of the model, such as LoRA modules, adapter layers, or task-specific heads; it is not the adapter rank alone. Frozen backbone parameters are excluded. A tied trainable block can be counted once in an explicit shared-coordinate parameterization, but the associated cover must be established for that parameterization. The independent-subtask formulas require separate trainable blocks as discussed above.

Radius parameters B_i and B. The constants B_i and B capture the radius of the feasible parameter spaces for each agent. Practical controls that motivate considering restricted parameter domains include:

- **weight decay** or ℓ_2 regularization,
- **gradient clipping**, which prevents excessively large updates,
- **KL-regularization** toward a reference model (as used in PPO, GRPO, DPO, and RLHF pipelines),

These controls do not by themselves guarantee a compact feasible parameter domain. The analysis separately assumes a closed, norm-bounded set, which can be enforced by explicit parameter constraints.

Lipschitz constants L_{step} and L_{seq}. These constants encode the smoothness of the policy with respect to its parameters at the token level and sequence level, respectively. Practical training and sensitivity controls include:

- **gradient clipping norms**,
- **normalization layers** (LayerNorm, RMSNorm),
- **temperature** and other sampling controls,
- **learning-rate schedules** and optimizer hyperparameters.

Gradient clipping and learning-rate schedules control optimization updates, not the uniform parameter sensitivity assumed in the theorems. Architecture and sampling controls can influence that sensitivity. Empirical perturbation analysis can provide local evidence, but does not certify a Lipschitz bound uniform over all feasible parameters, prompts, and histories.

Sequence length T. The parameter T bounds the number of decoding or reasoning steps within an agent's subtask. In practice, this bound depends on context-window limits, the decoding cap, and subtask structure, such as planner output length, reasoning-chain length, or solution depth.

Table 2: Summary of notation used throughout the paper.

Symbol	Meaning
$\mathbf{x} \in \mathcal{X}$	Input prompt sampled from distribution $\mathbb{D}$.
$\mathbf{y} = (a_1, \ldots, a_T)$	Generated output sequence from an LLM policy.
$\mathbf{y}^{(i)}$	Segment i of the decomposed output.
$\mathbf{y}^{(<i)}$	Concatenation of segments 1 to $i-1$ (previous agent outputs).
π_θ	SARL policy parameterized by $\theta \in \Theta$.
π_{θ_i}	MARL agent-i policy with parameters $\theta_i \in \Theta_i$.
$\theta = (\theta_1, \ldots, \theta_K)$	Joint MARL parameter vector.
d	Number of free trainable coordinates in $\Theta \subset \mathbb{R}^d$.
d_i	Number of free trainable coordinates in $\Theta_i \subset \mathbb{R}^{d_i}$.
B, B_i	Radius bounds for parameter sets $\|\theta\|_2 \le B$, $\|\theta_i\|_2 \le B_i$.
$T_{\max}$	Maximum generation length in SARL .
$T_{\max,i}$	Maximum generation length for agent i in MARL.
L_{step}	Lipschitz constant of token-level policy w.r.t. parameters .
$L_{\text{seq}} = T_{\max} L_{\text{step}}$	Sequence-level Lipschitz constant for SARL.
$L_{\text{seq},i} = T_{\max,i} L_{\text{step}}$	Sequence-level Lipschitz constant for agent i (MARL).
$R(x, y)$	Unified SARL reward for a full sequence (bounded in $[0, 1]$).
$r_i(x, y^{(i)}, y^{(<i)})$	MARL segment reward under *dependent* decomposition (Eq. (2.1)).
$r_i(x, y^{(i)})$	MARL segment reward under *independent* decomposition (Eq. (2.1)).
$R_{\text{dep}}(x, y)$	Decomposed dependent reward (Eq. (2.1)).
$R_{\text{indep}}(x, y)$	Decomposed independent reward (Eq. (2.1)).
$J(\pi)$	Population reward objective $\mathbb{E}_{x,y}[R(x, y)]$ (SARL).
$J_{\text{MARL}}(\pi)$	MARL population objective using $R_{\text{dep/indep}}$.
$\widehat{J}_n(\pi)$	Average conditional expected reward over n sampled prompts.
$N_{\text{SARL}}(\varepsilon, \delta)$	SARL prompt sample complexity, upper-bounded by $\lceil U_{\text{S}} \rceil$.
$N_{\text{MARL}}(\varepsilon, \delta)$	MARL prompt sample complexity, upper-bounded by $\lceil U_{\text{D}} \rceil$ or $\lceil U_{\text{I}} \rceil$ in the respective regimes.
K	Number of agents (segments) in MARL.

J Learning Errors in Stagewise Training

The stagewise guarantees combine an error-propagation inequality with assumptions on local learning rates. The reference policy π^0 is fixed independently of the training data. Write $Z_i = (X, Y^{(<i)})$ for the prompt and entering history together, and let d_i^0 be its distribution under the reference policy. The error $e_i = e_i(n)$ is defined by (3.2). Learned policies may depend on the training data and algorithmic randomness.

J.1 Proof of Theorem 3.8

Proof. Condition on the training data and algorithmic randomness, so that $\widehat{\pi}$ is fixed. Let $P^{(i)}$ be the transcript law that uses reference policies at stages $1, \dots, i$ and learned policies at stages $i+1, \dots, K$. Thus $P^{(0)} = P_{\widehat{\pi}}$ and $P^{(K)} = P_{\pi^0}$, with the same prompt distribution throughout. Adjacent hybrids $P^{(i)}$ and $P^{(i-1)}$ have entering-history law d_i^0 at stage i. Their rewards before stage i have identical expectations.

For a fixed prompt–history pair z and segment output y, let

$$q_i(z, y) := \mathbb{E}_{\widehat{\pi}_{i+1:K}} \left[\frac{1}{K} \sum_{j=i}^{K} r_j(X, Y^{(\le j)}) \,\middle|\, Z_i = z, Y^{(i)} = y \right].$$

The two hybrids use the same learned suffix in this definition, and $0 \le q_i \le w_i = (K - i + 1)/K$. The expectation difference of a function with range $[0, w_i]$ is at most w_i times total variation. Consequently,

$$|\mathbb{E}_{P^{(i)}} R - \mathbb{E}_{P^{(i-1)}} R| \le w_i \, \mathbb{E}_{Z_i \sim d_i^0} \mathrm{TV}\big(\pi_i^0(\cdot \mid Z_i), \widehat{\pi}_i(\cdot \mid Z_i)\big) = w_i e_i.$$

Summing over i proves the weighted bound. Both complete returns belong to $[0, 1]$, so the absolute difference between their expectations is also at most one, giving the stated minimum.

For history-independent policies and local rewards, write $J(\pi) = K^{-1} \sum_i \mathbb{E}_X \mathbb{E}_{\pi_i(\cdot|X)} r_i(X, Y^{(i)})$. Applying the bounded-reward TV inequality separately to these terms gives $|J(\widehat{\pi}) - J(\pi^0)| \le K^{-1} \sum_i e_i$. No independence between the agents' training datasets is needed for either deterministic inequality. □

The hybrid argument accounts for deployment histories that differ from d_i^0. It requires errors under the stated reference histories; errors measured under an unrelated training distribution need an additional coverage or transfer argument. No Lipschitz assumption on histories or parameters is needed. A small immediate local-reward loss alone does not imply a small conditional-policy TV error.

J.2 Learning rates and matched comparisons

Proof. For Corollary 3.1, fix $n \ge 1$. On the assumed simultaneous event in (3.4), Theorem 3.8 gives

$$J(\pi^0) - J(\widehat{\pi}) \le \sum_i w_i \left(b_i + \sqrt{C_i / n} \right) = B_w + \frac{A_w}{\sqrt{n}}.$$

For $\varepsilon > B_w$, solving $A_w/\sqrt{n} \le \varepsilon - B_w$ gives the integer budget in (3.5), with $n \ge 1$. If $J^\star$ is the optimum of a specified comparison class, let $g_0 = \max\{0, J^\star - J(\pi^0)\}$. Then $J^\star - J(\widehat{\pi}) \le g_0 + B_w + A_w/\sqrt{n}$, so the residual tolerance becomes $\varepsilon - g_0 - B_w > 0$. The gap thus accounts for a reference that is not class-optimal. □

The rate assumption concerns a specified learner and information model. It is not a convergence guarantee for Adam, PPO, or greedy local-reward training. Separate stage guarantees with failure probabilities δ_i yield the simultaneous event when $\sum_i \delta_i \le \delta$; their data may be correlated. The constants can incorporate this confidence allocation, but must satisfy the stated rate at the sample size being used.

Other learning rates. If a learner instead establishes $e_i(n) \le b_i + C_i/n$ simultaneously, the same proof gives $J^\star - J(\widehat{\pi}) \le g_0 + B_w + D_w/n$, where $D_w = \sum_i w_i C_i$. A sufficient integer budget is $\max\{1, \lceil D_w/(\varepsilon - g_0 - B_w)\rceil\}$ when the denominator is positive. The sample dependence therefore follows from the proved local learning rate; it cannot be obtained merely by squaring a stagewise accuracy requirement.

Comparison with a single-agent learner. Suppose a SARL learner has a guarantee $J^\star - J(\widehat{\pi}_{\mathrm{S}}) \le g_{\mathrm{S}} + b_{\mathrm{S}} + \sqrt{C_{\mathrm{S}}/n}$, while the modular learner has the guarantee above with gap g_{M}. The comparison uses the same evaluation reward and optimum $J^\star$, with both reference policies and any difference in their values specified. A common reference gives equal gaps. Write

$$\rho_{\mathrm{M}} = \varepsilon - g_{\mathrm{M}} - B_w > 0, \qquad \rho_{\mathrm{S}} = \varepsilon - g_{\mathrm{S}} - b_{\mathrm{S}} > 0.$$

Their real-valued sufficient thresholds are

$$V_{\mathrm{M}} = A_w^2/\rho_{\mathrm{M}}^2, \qquad V_{\mathrm{S}} = C_{\mathrm{S}}/\rho_{\mathrm{S}}^2.$$

Each threshold gives the integer budget $n \ge \max\{1, \lceil V \rceil\}$. Thus $A_w^2/\rho_{\mathrm{M}}^2 \le C_{\mathrm{S}}/\rho_{\mathrm{S}}^2$ suffices to order these budgets at matched confidence levels and observation costs. A simultaneous probability statement also requires allocating the total failure probability across both guarantees. Ordering sufficient budgets alone does not order the learners' actual minimum sample requirements.

For zero bias terms and a common reference gap, the residual tolerances agree. If $C_i = C_0 > 0$, the dependent and independent propagation coefficients satisfy

$$A_{\mathrm{D}}^2 = \frac{(K+1)^2}{4} C_0, \qquad A_{\mathrm{I}}^2 = C_0.$$

Their real-valued thresholds have ratio $(K+1)^2/4$. In contrast, if a justified capacity match gives $C_i = C_{\mathrm{S}}/K$ with $C_{\mathrm{S}} > 0$,

$$\frac{A_{\mathrm{D}}^2}{C_{\mathrm{S}}} = \frac{(K+1)^2}{4K}, \qquad \frac{A_{\mathrm{I}}^2}{C_{\mathrm{S}}} = \frac{1}{K}.$$

These are conditional comparisons of the displayed certificates. A parameter count alone does not establish either relation between the learning constants. Here n is the number of training observations available to each stage. If one shared reference trajectory supplies all K stage examples, n such trajectories yield Kn local examples. With distinct stage datasets of size n, the total is Kn observations. SARL comparisons must use the same resource.

J.3 When joint analysis is stronger

Proposition J.1 (A conditional KL alternative)**.** *For fixed learned policies, define*

$$\kappa_i := \mathbb{E}_{Z_i \sim d_i^0} D_{\mathrm{KL}}\big(\pi_i^0(\cdot \mid Z_i) \,\|\, \widehat{\pi}_i(\cdot \mid Z_i)\big)\,.$$

For any common reward $R \in [0,1]$, $|J(\pi^0) - J(\widehat{\pi})| \le \min\{1, \sqrt{\frac{1}{2}\sum_i \kappa_i}\}$. For the normalized prefix reward, the weighted bound $\sum_i w_i \sqrt{\kappa_i/2}$ also holds, so one may take the smaller bound.

Proof. The KL chain rule for the joint prompt–transcript laws gives $D_{\mathrm{KL}}(P_{\pi^0} \| P_{\widehat{\pi}}) = \sum_i \kappa_i$ because the prompt law is common. Pinsker's inequality and the bounded-reward TV inequality give the first claim. Applying Pinsker conditionally and then Jensen's inequality gives $e_i \le \sqrt{\kappa_i/2}$, proving the second claim. If a KL term is infinite the corresponding bound is uninformative. □

In particular, if $\kappa_i \le C_i/n$, the joint bound is $\sqrt{\sum_i C_i/(2n)}$, which can be much smaller than the sum of local square roots. An $O(n^{-1/2})$ excess-log-loss guarantee relative to the reference conditional distribution instead gives only $O(n^{-1/4})$ TV control through Pinsker; a C_i/n KL rate requires its own justification. Appendix J.5 gives a finite-class training example with a complementary product-class analysis. Separate conditional maximum likelihood over a Cartesian policy class also maximizes the joint trajectory likelihood, since its log likelihood is a sum of stage terms. Stagewise training alone therefore does not imply an unavoidable quadratic sample penalty.

J.4 IMPERFECT JOINT EMPIRICAL OPTIMIZATION

Proposition J.2 (Joint training errors)**.** *Fix a policy class and let J, $\widehat{J}_n$, and $\widetilde{J}_n$ denote its population objective, exact conditional-value empirical objective, and estimated empirical objective. Let $\varepsilon_{\mathrm{stat}} > 0$ and $\nu, \tau \geq 0$. Suppose events of failure probabilities at most δ_{s}, δ_{v}, and δ_{o} respectively guarantee*

$$\sup_{\pi} |J(\pi) - \widehat{J}_n(\pi)| \leq \varepsilon_{\mathrm{stat}}/2,$$
$$\sup_{\pi} |\widetilde{J}_n(\pi) - \widehat{J}_n(\pi)| \leq \nu,$$
$$\widetilde{J}_n(\widehat{\pi}) \geq \sup_{\pi} \widetilde{J}_n(\pi) - \tau.$$

Then, with probability at least $1 - \delta_{\mathrm{s}} - \delta_{\mathrm{v}} - \delta_{\mathrm{o}}$,

$$\sup_{\pi} J(\pi) - J(\widehat{\pi}) \leq \varepsilon_{\mathrm{stat}} + 2\nu + \tau.$$

Proof. On the intersection of the three events,

$$\begin{aligned}\sup_{\pi} J(\pi) - J(\widehat{\pi}) &\leq \varepsilon_{\mathrm{stat}} + \sup_{\pi} \widehat{J}_n(\pi) - \widehat{J}_n(\widehat{\pi}) \\ &\leq \varepsilon_{\mathrm{stat}} + 2\nu + \sup_{\pi} \widetilde{J}_n(\pi) - \widetilde{J}_n(\widehat{\pi}) \\ &\leq \varepsilon_{\mathrm{stat}} + 2\nu + \tau.\end{aligned}$$

A union bound gives the stated confidence; event independence is unnecessary. □

For a target $\varepsilon > 2\nu + \tau$, the uniform-convergence guarantees apply with statistical tolerance $\varepsilon - 2\nu - \tau$ and failure probability δ_{s}. Deterministically guaranteed tolerances have zero corresponding failure probability. Bounds on ν and τ must be established for the chosen rollout estimator and optimizer. The proposition concerns a joint objective, rather than greedy local-reward training.

J.5 A FINITE-CLASS ILLUSTRATION WITH REFERENCE DEMONSTRATIONS

We instantiate the local-error condition with a realizable finite-class supervised learner trained on reference demonstrations. Compounding imitation errors under reference-state training are established phenomena (Ross & Bagnell, 2010; Ross et al., 2011). The finite-class setting gives explicit rates under specified realizability and sampling assumptions.

Training algorithm. Let $\pi^0 = (\pi^0_1, \ldots, \pi^0_K)$ be a deterministic reference policy and $\mathcal{H}_i$ a finite class of deterministic stage policies, with $|\mathcal{H}_i| = M_i$ and $\pi^0_i \in \mathcal{H}_i$. Observe n complete reference demonstrations (X_j, Y^0_j), where the prompts are iid from $\mathbb{D}$ and $Y^0_{j,i} = \pi^0_i(X_j, Y^0_{j,<i})$. For stages $i = 1, \ldots, K$, choose any $\widehat{\pi}_i \in \mathcal{H}_i$ satisfying $\widehat{\pi}_i(X_j, Y^0_{j,<i}) = Y^0_{j,i}$ for every j, then freeze it. At deployment, use $\widehat{Y}_i = \widehat{\pi}_i(X, \widehat{Y}_{<i})$. Thus training uses reference histories, while deployment uses generated histories. Each demonstration supplies K labels: there are n independent trajectories and nK stage labels, not nK independent trajectories.

Proposition J.3 (Consistent finite-class stage learners)**.** *For $0 < \delta < 1$, define*

$$e_i := \Pr_X\left[\widehat{\pi}_i(X, Y^0_{<i}) \neq Y^0_i\right], \qquad c_i := \log M_i + \log(K/\delta), \qquad w_i := \frac{K - i + 1}{K}.$$

With probability at least $1 - \delta$ over the demonstrations, simultaneously $e_i \leq \min\{1, c_i/n\}$ for all i. The normalized prefix-mismatch risk

$$R_{\mathrm{pre}}(\widehat{\pi}) := \mathbb{E}_X\left[\frac{1}{K}\sum_{j=1}^{K} \mathbf{1}\{\widehat{Y}_{1:j} \neq Y^0_{1:j}\}\right]$$

satisfies

$$J_R(\pi^0) - J_R(\widehat{\pi}) \leq R_{\mathrm{pre}}(\widehat{\pi}) \leq \min\left\{1, \sum_i w_i e_i\right\} \leq \min\left\{1, \frac{\sum_i w_i c_i}{n}\right\}$$

for any reward $R = K^{-1}\sum_j r_j(X, Y_{1:j})$ with $0 \le r_j \le 1$. Consequently, $n \ge \lceil \sum_i w_i c_i/\varepsilon \rceil$ suffices for reward loss relative to π^0 at most ε. If $c_i \le c$, this budget is $\lceil (K+1)c/(2\varepsilon) \rceil$. For history-independent policies and an average of bounded local rewards $r_i(X, Y_i)$, the corresponding guarantee uses $K^{-1}\sum_i e_i$ and the sufficient budget $\lceil c/\varepsilon \rceil$.

Proof. For a fixed stage-i hypothesis with reference error greater than η, the probability of consistency on all n demonstrations is at most $(1-\eta)^n \le e^{-n\eta}$. Union bounds over M_i hypotheses and then K stages, using $\eta = c_i/n$, give the simultaneous error bounds; when $c_i/n \ge 1$ the bound holds automatically. Examples at a fixed stage are iid across demonstrations. Independence between stages is not needed.

Couple the reference and learner on the same fresh prompt. Before their first disagreement, their histories coincide, so disagreement by stage j has probability at most $\sum_{i \le j} e_i$. Averaging over j gives the weights w_i in Theorem 3.8. A bounded prefix reward can differ only after such a disagreement, proving the reward inequality. For history-independent policies and local rewards, each component depends only on its own disagreement. Finally, $\sum_i w_i = (K+1)/2$ gives the stated homogeneous budgets. □

A direct product-class consistency bound. Induction over the stages shows that the learner reproduces every training trajectory exactly. Its complete policy lies in the fixed product class $\prod_i \mathcal{H}_i$, of size $\prod_i M_i$. For any fixed member with $R_{\mathrm{pre}} > \eta$, the probability of zero prefix loss on n independent demonstrations is at most $e^{-n\eta}$: for a bounded loss $\ell \in [0,1]$, $\Pr(\ell = 0) \le 1 - \mathbb{E}\ell$. Thus a separate union bound gives

$$\Pr\left[R_{\mathrm{pre}}(\widehat{\pi}) \le \min\left\{ 1, \frac{\sum_i \log M_i + \log(1/\delta)}{n} \right\} \right] \ge 1 - \delta.$$

This bound also controls reward loss. The realizable finite-class learner has a $1/n$ rate and no intrinsic K^2 sample penalty. A joint consistent learner in the same full product class has this same guarantee; the selected policies need not coincide. Comparisons with SARL must retain the same demonstrations, evaluation reward, and comparator. In particular, evaluating independent predictions by prefix correctness still propagates prefix failures, so changing the reward cannot establish an architectural separation.

J.6 Illustrative teacher-forced linear regression

We analyze teacher-forced linear regression for a variant of the synthetic task. The criterion is expected excess mean squared error (MSE), rather than bounded-reward PAC performance. The guarantees concern the specified regression learner; other estimators require their own error analysis.

Let $X = (x_1, \dots, x_K)$ and consider the linear structural model

$$Y = CY + f(X) + \xi, \qquad f_i(X) = w_i^\top x_i, \qquad \mathbb{E}[\xi \mid X] = 0,$$

with finite second moments. The strictly lower-triangular matrix C is *known and fixed*, so $G := (I - C)^{-1}$ exists and $m(X) := \mathbb{E}[Y \mid X] = Gf(X)$. Given n independent training pairs, form each local regression target using the observed training history:

$$Z_i := Y_i - \sum_{j<i} C_{ij} Y_j = f_i(X) + \xi_i.$$

For full-column-rank local designs, fit $\widehat{w}_i$ by ordinary least squares on (x_i, Z_i), set $\widehat{f}_i(X) = \widehat{w}_i^\top x_i$, and freeze the fitted regressors. At test time, use predicted rather than observed histories: $\widehat{m}_i(X) = \widehat{f}_i(X) + \sum_{j<i} C_{ij}\widehat{m}_j(X)$, hence $\widehat{m}(X) = G\widehat{f}(X)$. Assume the prediction errors have finite second moments. The test pair is independent of the training sample.

Proposition J.4 (Propagation of local regression errors)**.** *Let $e(X) = \widehat{f}(X) - f(X)$ and $Q = \mathbb{E}[e(X)e(X)^\top]$, with expectation over training and a fresh test input. The expected excess MSE relative to the conditional mean is*

$$\begin{aligned} \mathcal{E}_K &:= \frac{1}{K}\mathbb{E}\big[\|Y - \widehat{m}(X)\|_2^2 - \|Y - m(X)\|_2^2\big] \\ &= \frac{1}{K}\mathbb{E}\|Ge(X)\|_2^2 = \frac{1}{K}\operatorname{tr}(GQG^\top). \end{aligned} \tag{J.1}$$

For the averaging coupling $C_{ij} = 1/(i-1)$ for $i > j$ and zero otherwise, $G_{ii} = 1$ and $G_{ij} = 1/j$ for $i > j$. If additionally $Q = vI$, then

$$\mathcal{E}_K = vA_K, \qquad A_K := 1 + \frac{1}{K}\sum_{j=1}^{K-1}\frac{K-j}{j^2} < 1 + \frac{\pi^2}{6}, \qquad K \geq 1. \tag{J.2}$$

Proof. The two recursions imply $\widehat{m} - m = Ge$. Because the test innovations have conditional mean zero and the training sample is independent of the test pair, the cross term between $Ge(X)$ and $G\xi$ has expectation zero. Expanding the two squared losses proves (J.1). To obtain the inverse for averaging coupling, set $G_{ii} = 1$, $G_{ij} = 1/j$ for $i > j$, and $G_{ij} = 0$ for $i < j$. For $i > j$,

$$(CG)_{ij} = \frac{1}{i-1}\left(1 + \sum_{h=j+1}^{i-1}\frac{1}{j}\right) = \frac{1}{j}.$$

The diagonal and upper-triangular entries also give $G = I + CG$, proving the inverse formula. Under $Q = vI$, the trace equals $v\|G\|_F^2$. Column j has one diagonal entry and $K - j$ entries of size $1/j$, which gives A_K. Finally, $(K-j)/K < 1$ and $\sum_{j=1}^{\infty} j^{-2} = \pi^2/6$ prove the bound, including $K = 1$. □

The isotropic condition $Q = vI$ requires assumptions beyond sequential execution. A sufficient special case is correctly specified OLS with independent stagewise design/noise pairs, identical feature dimensions and feature/noise laws, and finite prediction-error second moments. Conditional unbiasedness and the identical local estimation problems then give zero cross moments and equal variances. Otherwise, the full matrix Q is needed; if $\mu = \mathbb{E}e \neq 0$, using only $\mathrm{Cov}(e)$ would omit $\|G\mu\|_2^2/K$. In particular, correlated or unequal stage errors can produce a different amplification factor.

Writing $\Omega(X) = \mathbb{E}[\xi\xi^\top \mid X]$, total expected test MSE also includes the irreducible term $K^{-1}\mathbb{E}_X \operatorname{tr}(G\Omega(X)G^\top)$. If the design and moment assumptions and the rate $Q = v(n)I$, $v(n) \leq c/n$ hold for $n \geq n_0 \geq 1$, then $\mathcal{E}_K \leq A_K c/n$; thus $n \geq \max\{n_0, \lceil A_K c/\varepsilon \rceil\}$ suffices for expected excess MSE at most ε. This is an expectation guarantee, not a high-probability PAC claim. Exact OLS models finite-data estimation error; finite optimization iterations require their own error analysis. Estimated coupling coefficients would likewise introduce additional terms. Under the stated estimation and moment conditions, averaging dependence need not create a K^2 penalty. This conclusion concerns the regression model and does not establish an architectural advantage.

J.7 Exact Gaussian learning rates and a matched joint comparator

Proposition 3.9 specializes the preceding model so that the local second moments follow from a specified estimator and data law, rather than a learning-rate premise. The local Gaussian OLS risk is standard (Belkin et al., 2020); the calculation gives explicit dependence on dimension and sample size. Throughout, $p \geq 1$, $\sigma^2 > 0$, $n > p + 1$, and C and σ^2 are known. For every independent observation, the K inputs $x_i \sim \mathcal{N}(0, I_p)$ and innovations $\xi_i \sim \mathcal{N}(0, \sigma^2)$ are mutually independent. The vectors $w_i \in \mathbb{R}^p$ are unrestricted deterministic parameters. Each observation contains the entire pair (X, Y), so both learners have the same data.

Proof of Proposition 3.9. **Local estimation.** Let $D_i \in \mathbb{R}^{n\times p}$ be the design matrix for stage i and $\boldsymbol{\xi}_i$ its vector of training innovations. Removing the known coupling gives $\boldsymbol{Z}_i = D_i w_i + \boldsymbol{\xi}_i$. Gaussian designs have full column rank almost surely, and

$$\widehat{w}_i - w_i = (D_i^\top D_i)^{-1} D_i^\top \boldsymbol{\xi}_i, \qquad \mathbb{E}[(\widehat{w}_i - w_i)(\widehat{w}_i - w_i)^\top \mid D_i] = \sigma^2 (D_i^\top D_i)^{-1}.$$

For an independent test input, $e_i = x_i^\top(\widehat{w}_i - w_i)$ therefore has second moment $\sigma^2 \mathbb{E}\operatorname{tr}(D_i^\top D_i)^{-1}$. The jth diagonal entry of the inverse Gram matrix is $[d_j^\top (I - P_{-j}) d_j]^{-1}$, where d_j is the jth column and P_{-j} projects onto the other $p-1$ columns. Conditional on those columns, the denominator has a χ^2_{n-p+1} distribution; its reciprocal mean is $1/(n-p-1)$. Thus

$$Q = v_n I, \qquad v_n = \frac{\sigma^2 p}{n-p-1}. \tag{J.3}$$

The off-diagonal moments vanish because the local design/noise pairs and test inputs are independent and each error is centered. Substituting into Proposition J.4 gives $\mathcal{E}_K = v_n \mathcal{A}_K(C)$.

Matched joint estimation. The joint comparator maximizes the correctly specified conditional density of the complete vector Y given X, over the same tuple $(w_1, \dots, w_K)$. The change of variables $(I - C)Y = f(X) + \xi$ has unit Jacobian determinant. Consequently, up to a parameter-independent constant, its negative log-likelihood on the sample is

$$\frac{1}{2\sigma^2} \sum_{\ell=1}^{n} \sum_{i=1}^{K} \left(Y_i^{(\ell)} - \sum_{j<i} C_{ij} Y_j^{(\ell)} - w_i^\top x_i^{(\ell)} \right)^2 .$$

This objective separates into exactly the K local least-squares problems. Their almost surely unique solutions are identical to the stagewise fits, and both predictors use $\widehat{m} = G\widehat{f}$. This proves equality of the estimators and their expected excess risks. The joint comparator uses the known covariance $\sigma^2 GG^\top$ of $Y \mid X$; unweighted output-space least squares would be a different objective.

Same graph, different amplification. For $a = 1$, Proposition J.4 gives $G_{ij} = 1/j$ for $i > j$ and the stated bounded factor. For $a = 2$, let $G_{ii} = 1$ and $G_{ij} = 2i/[j(j+1)]$ for $i > j$, with zero entries above the diagonal. For $i > j$,

$$(C^{(2)}G)_{ij} = \frac{2}{i-1} \left(1 + \sum_{h=j+1}^{i-1} \frac{2h}{j(j+1)} \right) = \frac{2i}{j(j+1)} = G_{ij}.$$

The diagonal entries of $C^{(2)}G$ vanish, so $G = I + C^{(2)}G$ proves the inverse formula. Summing the squared entries gives the $C^{(2)}$ expression in (3.7) exactly. Moreover,

$$\lim_{K\to\infty} \frac{\mathcal{A}_K(C^{(2)})}{K^2} = \frac{4}{3} \sum_{j=1}^{\infty} \frac{1}{j^2(j+1)^2} = \frac{4}{3} \left(\frac{\pi^2}{3} - 3 \right) > 0.$$

For fixed j, $K^{-3} \sum_{i=j+1}^{K} i^2 \to 1/3$; these terms are bounded by 1, so the summable sequence $1/[j^2(j+1)^2]$ justifies passing the limit through the sum. The final identity follows from $1/[j^2(j+1)^2] = 1/j^2 + 1/(j+1)^2 - 2/[j(j+1)]$. The positive limit establishes quadratic growth and identifies its leading constant. □

Exact expected-error budgets and interpretation. For either estimator, fixed C, and $\varepsilon > 0$, the smallest integer sample size within the stated regime $n > p + 1$ satisfying $\mathcal{E}_K \le \varepsilon$ is

$$n_{\min}(\varepsilon; C) = p + 1 + \left\lceil \frac{\sigma^2 p \mathcal{A}_K(C)}{\varepsilon} \right\rceil . \tag{J.4}$$

This follows from an exact risk identity, rather than a comparison of potentially loose upper bounds. For fixed p, σ^2, and ε, averaging has a budget bounded uniformly in K, whereas doubled averaging has a budget of order K^2. The dimensions are matched between the two couplings and the two learners at each K, but the total trainable dimension is Kp; these statements do not hold total dimension fixed while increasing K.

The irreducible normalized MSE is $\mathcal{R}_K^\star = \sigma^2 \mathcal{A}_K(C)$, hence $\mathcal{E}_K / \mathcal{R}_K^\star = p/(n-p-1)$ and total MSE is $\sigma^2 \mathcal{A}_K(C)[1 + p/(n-p-1)]$. Coupling amplifies innovations as well as estimation errors. The matched joint and stagewise estimators have equal risk, so this amplification reflects task sensitivity shared by both, without an architectural advantage or disadvantage. Joint likelihood maximization here is distinct from the complete-return empirical maximization in Theorem 3.2. Known coupling, Gaussian features, independent innovations, and exact least squares are essential to this comparison; the calculation does not provide a neural optimization rate or an unknown-coupling guarantee. The $a = 2$ case lies outside the coupling range examined in the empirical study.

J.8 When feature restriction improves expected risk

The matched-class comparison above isolates sensitivity shared by joint and stagewise fitting. We now compare two nested feature classes to determine when a reduction in estimation error outweighs

the loss of expressivity. Both learners receive the same complete observations and are evaluated against the same conditional mean $m(X) = Gf(X)$.

Use the Gaussian model of Proposition 3.9, with $p \geq 2$. Before observing the training sample, fix a subset $S_i \subset \{1, \ldots, p\}$ of size q, where $1 \leq q < p$, for each stage. The restricted learner regresses Z_i on x_{i,S_i}; the full learner regresses Z_i on all p features. Both use ordinary least squares without an intercept and predict recursively using the known C. Write $\widehat{m}_{\mathrm{R}} = G\widehat{f}_{\mathrm{R}}$ and $\widehat{m}_{\mathrm{F}} = G\widehat{f}_{\mathrm{F}}$ for their predictions. Define the omitted signal and its propagation weight by

$$b_i := \|w_{i,S_i^c}\|_2^2, \qquad s_i := \frac{\|G_{:,i}\|_2^2}{K}, \qquad \bar{b}_G := \frac{\sum_{i=1}^K s_i b_i}{\mathcal{A}_K(C)}, \qquad \sum_{i=1}^K s_i = \mathcal{A}_K(C). \tag{J.5}$$

Here S_i^c is the complement of S_i and $G_{:,i}$ is column i of G. Thus $\bar{b}_G$ averages the omitted signal according to its effect on the complete prediction.

Proposition J.5 (Exact comparison of nested feature classes)**.** *Under these assumptions, for $n > p+1$, the normalized expected excess MSEs are*

$$\begin{aligned} \mathcal{E}_{\mathrm{R}} &= \sum_{i=1}^K s_i \left[b_i + \frac{q(\sigma^2 + b_i)}{n-q-1} \right] = \mathcal{A}_K(C) \left[\bar{b}_G + \frac{q(\sigma^2 + \bar{b}_G)}{n-q-1} \right], \\ \mathcal{E}_{\mathrm{F}} &= \mathcal{A}_K(C) \frac{\sigma^2 p}{n-p-1}. \end{aligned} \tag{J.6}$$

Consequently,

$$\mathcal{E}_{\mathrm{R}} \leq \mathcal{E}_{\mathrm{F}} \quad \Longleftrightarrow \quad \bar{b}_G(n-p-1) \leq \sigma^2(p-q). \tag{J.7}$$

If $\bar{b}_G > 0$, the restricted learner has no larger risk exactly when

$$n \leq n_\star(C) := p + 1 + \frac{\sigma^2(p-q)}{\bar{b}_G}, \qquad n > p+1. \tag{J.8}$$

If $\bar{b}_G = 0$, its risk is strictly smaller for every $n \geq p+1$. The best predictor in the restricted product linear class has excess MSE $\sum_i s_i b_i = \mathcal{A}_K(C)\bar{b}_G$, which is also the limit of $\mathcal{E}_{\mathrm{R}}$ as $n \to \infty$.

Proof. For stage i, partition the training design into its included and omitted columns. The residualized target satisfies

$$\boldsymbol{Z}_i = D_{i,S_i} w_{i,S_i} + \boldsymbol{\eta}_i, \qquad \boldsymbol{\eta}_i = D_{i,S_i^c} w_{i,S_i^c} + \boldsymbol{\xi}_i.$$

The effective noise $\boldsymbol{\eta}_i$ is independent of D_{i,S_i} and has distribution $\mathcal{N}(0, (\sigma^2 + b_i) I_n)$. The Gaussian OLS calculation used in (J.3), now with q features, therefore gives

$$\mathbb{E}\|\widehat{w}_{\mathrm{R},i} - w_{i,S_i}\|_2^2 = \frac{q(\sigma^2 + b_i)}{n-q-1}.$$

For a fresh test input, the local prediction error is

$$e_{\mathrm{R},i} = x_{i,S_i}^\top (\widehat{w}_{\mathrm{R},i} - w_{i,S_i}) - x_{i,S_i^c}^\top w_{i,S_i^c}, \qquad \mathbb{E}e_{\mathrm{R},i}^2 = b_i + \frac{q(\sigma^2 + b_i)}{n-q-1}.$$

The two terms have zero cross moment because the fresh included and omitted features are independent and centered. Conditional on the training sample, local errors at different stages also have zero cross moments, by the independent centered test features. Thus the error second-moment matrix is diagonal. Proposition J.4 yields the restricted risk in (J.6); the full risk follows from Proposition 3.9. The omitted-feature calculation is the standard Gaussian regression risk formula (Belkin et al., 2020).

Subtracting the two risks gives

$$\mathcal{E}_{\mathrm{R}} - \mathcal{E}_{\mathrm{F}} = \frac{\mathcal{A}_K(C)(n-1)}{(n-q-1)(n-p-1)} \left[\bar{b}_G(n-p-1) - \sigma^2(p-q) \right].$$

The prefactor is positive in the stated regime, proving the comparison and threshold, including the strict inequality when $\bar{b}_G = 0$. Finally, for arbitrary fixed local coefficients θ_i in the restricted class, the same independence calculation gives excess MSE $\sum_i s_i[\|\theta_i - w_{i,S_i}\|_2^2 + b_i]$. It is minimized at $\theta_i = w_{i,S_i}$, proving the approximation floor. □

Table 3: An exact change in the preferred feature class with coupling strength. Both cases have the same two-stage dependency graph, feature classes, and sample size $n = 8$. Entries are normalized expected excess MSE; smaller values are better.

Coupling c	Restricted class	Full class	Lower risk
1	$\mathbf{0.5475}$	0.6000	Restricted
2	1.2350	$\mathbf{1.2000}$	Full

How propagation changes the choice. The threshold compares the dimension reduction $p - q$ with the omitted signal weighted by downstream sensitivity. When $\bar{b}_G > 0$, restriction helps at sample sizes below the threshold and hurts above it; equality holds at the threshold if it is an integer. All comparisons are within the common finite-risk regime $n \geq p + 2$, so the restricted learner never wins there if $n_\star(C) < p + 2$. If every b_i is equal, changing C rescales both risks by the same factor and does not change their ordering. With unequal omitted signals, changing the coupling can change $\bar{b}_G$ and reverse the preferred class.

A two-stage reversal on the same dependency graph. Let $p = 2$, $q = 1$, and consider

$$Y_1 = \beta_1 x_1 + \gamma_1 z_1 + \xi_1, \qquad Y_2 = cY_1 + \beta_2 x_2 + \gamma_2 z_2 + \xi_2, \tag{J.9}$$

where x_1, z_1, x_2, z_2 are independent standard Gaussians, the innovations are independent $\mathcal{N}(0, \sigma^2)$, and c is known. The full learner uses (x_i, z_i), while the restricted learner uses only x_i; both receive all features and both responses in every observation. Here $b_i = \gamma_i^2$, $s_1 = (1 + c^2)/2$, and $s_2 = 1/2$, so

$$\bar{b}_G = \frac{(1 + c^2)\gamma_1^2 + \gamma_2^2}{2 + c^2}, \qquad n_\star(c) = 3 + \frac{\sigma^2(2 + c^2)}{(1 + c^2)\gamma_1^2 + \gamma_2^2}.$$

For $\gamma_1^2 > \gamma_2^2$, stronger coupling increases $\bar{b}_G$ and lowers the threshold: omitting upstream information becomes more costly. Set $n = 8$, $\sigma^2 = 1$, $\gamma_1 = 1/2$, and $\gamma_2 = 1/10$, giving $b_1 = 1/4$ and $b_2 = 1/100$. The values in Table 3 follow directly from (J.6) and are exact expectations, not simulation estimates. The included coefficients β_1, β_2 can be any fixed values and do not affect these risks.

The reversal is not explained solely by an increase in overall noise scale. The common Bayes MSE is $\sigma^2 \mathcal{A}_2(C)$, equal to 1.5 for $c = 1$ and 3 for $c = 2$. Dividing by it gives restricted excess risks 0.365 and approximately 0.4117, respectively, while the full learner's ratio is 0.4 in both cases.

Relation to the comparative framework. This example combines finite-sample estimation, class approximation, and error propagation in an exact comparison. It shows why a smaller local dimension alone does not determine which class is preferable: the amount and location of omitted signal also matter. The threshold is a population comparison for specified classes and parameters; applying it as a data-driven selection rule would require estimating these quantities and accounting for selection error.

Within either feature class, maximizing the specified joint Gaussian trajectory likelihood with covariance $\sigma^2 GG^\top$ still separates into the same local least-squares fits. For the restricted class, this likelihood is generally misspecified conditional on the full input X, but the algebraic equivalence remains valid. Thus the reversal concerns class restriction, not an intrinsic ranking of single- and multi-agent architectures or of joint and stagewise optimization. The approximation floor here is measured in expected excess MSE and is distinct from the reward discrepancy α in the bounded-reward analysis.

K Extended Related Work

Empirical single-agent and multi-agent comparisons. Reported benefits of multi-agent LLM systems depend on training, task structure, and evaluation resources. Optima trains a shared model for collaborating agents and reports improvements in task performance and token efficiency (Chen et al., 2025). Kim et al. (2026) evaluate frozen-model agent architectures across tasks and develop a predictive account of coordination gains and costs. Tran & Kiela (2026) combine an information-theoretic perspective with budget-aware multi-hop reasoning evaluations, finding single-agent systems competitive with or better than several multi-agent configurations. These studies provide comparative insights across training and inference settings. Our analysis concerns how policy-class complexity, objective mismatch, and the learning procedure enter learning-sample guarantees for modular sequential policies. Conditional-value prompt budgets and reference-demonstration budgets are distinct from inference resources.

Structured RL and hierarchical comparisons. Factored models exploit environment structure (Kearns & Koller, 1999; Koller & Parr, 1999), while value-decomposition methods constrain how local utilities represent a team objective (Sunehag et al., 2017; Rashid et al., 2018). Robert et al. (2023) derive a lower bound for goal-conditioned hierarchical RL and compare hierarchical and monolithic learning through state, action, and temporal abstractions. Chen et al. (2021) derive episodic regret guarantees using factored transitions and rewards and compare them with unfactored MDP guarantees. Wang et al. (2021) analyze cooperative fitted Q-iteration, showing how value-function expressiveness and the training distribution affect convergence. These results quantify how environment structure, value representations, and training procedures affect RL learning. Our joint analysis studies policy-class generalization under conditional-value access and compares sufficient budgets relative to the stated class optima. Reward discrepancy and a separate class-optimum gap account for changes in the objective and attainable performance.

Sequential prediction and stagewise learning. Training on reference histories and executing on generated histories is a classical imitation-learning problem. Ross & Bagnell (2010) analyze compounding errors and forward training; Ross et al. (2011) address induced distribution shift through interactive data collection. We use this propagation principle for normalized segment rewards, separating position-dependent downstream sensitivity from the local learning-rate premise. That separation is necessary when comparing demonstration budgets: execution dependence alone does not specify how quickly the local errors decrease with data.

Comparative synthesis and illustrative calculations. These lines of work supply complementary tools with different information models, objectives, and comparators. Our framework combines them to compare SARL and MARL learning efficiency under explicit conditions. Gaussian least-squares risk calculations are established (Belkin et al., 2020); Proposition 3.9 uses them to isolate shared coupling sensitivity from differences between matched joint and stagewise learners. Each comparison uses the stated assumptions and error criterion of its setting.