Pareto-Optimal Few-Shot Adaptation in 2026: Cost-Regularized In-Context Learning + Retrieval + Low-Rank PEFT under Cross-Domain Shift

Liz Lemma Future Detective

January 20, 2026

Abstract

Few-shot learning (FSL) has historically been studied via episodic supervised tasks with gradient-based meta-learning, metric learning, or transfer learning. In the 2026 foundation-model era, adaptation is no longer a single knob: practitioners can use in-context learning (ICL) with retrieved demonstrations, parameter-efficient finetuning (PEFT) such as low-rank adapters, or full finetuning—each with distinct latency, energy, and reliability tradeoffs, especially under cross-domain shift. Inspired by recent survey taxonomies that unite classic FSL with emerging ICL and hybrid settings, we formalize a cost-regularized few-shot adaptation problem in which an algorithm chooses per task (i) which demonstrations to retrieve and place in context and (ii) whether to apply a constrained PEFT update. We analyze a clean special case—linear prediction in a pretrained embedding with task-dependent domain shift—where ICL corresponds to kernel/ridge regression in a fixed representation and PEFT corresponds to learning a low-rank correction to representation or readout. Our main results characterize the error–cost Pareto frontier: we give an explicit hybrid policy that is provably near-Pareto-optimal (up to logarithmic factors) over a broad class of adaptation policies, and we provide matching lower bounds showing when ICL-only must fail under shift. We outline how these guarantees motivate a practical gating algorithm and propose experiments on cross-domain FSL benchmarks (Meta-Dataset/Meta-Album/BSCD-FSL) and multimodal few-shot tasks, reporting cost-aware metrics (tokens, updated parameters, wall-clock) alongside accuracy.

Table of Contents

2. 2. Related Work and Positioning: meta-learning vs transfer learning debate; ICL theory (implicit GD/Bayes/linear regression); retrieval-augmented prompting; PEFT for few-shot; cross-domain FSL; connect explicitly to the survey's taxonomy and open problems (evaluation, domain shift, Green AI).

3. 3. Formal Problem: Cost-Regularized Few-Shot Adaptation (CR-FSA): define tasks, support/query, retrieval memory, adaptation actions (ICL/PEFT/hybrid), and cost model; define objectives (Lagrangian and constrained forms) and the Pareto frontier concept.

4. 4. Clean Special Case Model: Linearized Foundation Model Regime: task model $y = \langle w_T, A_T\phi(x)\rangle + \xi$, assumptions on ϕ, A_T, noise, and retrieval distribution; define shift measure and estimation setting; map ICL to ridge regression and PEFT to low-rank correction.

5. 5. Algorithms: (i) Analytic Hybrid Policy in the Special Case; (ii) Practical Gating Policy Template: how to estimate shift/uncertainty cheaply, select demonstrations, and decide whether to apply PEFT under a budget.

6. 6. Upper Bounds: Generalization and Cost–Error Tradeoffs: derive explicit excess risk decomposition (estimation + shift + approximation); prove near-Pareto optimality of the hybrid policy among admissible policy classes; discuss dependence on K, L, r, s, δ.

7. 7. Lower Bounds and Impossibility: show worst-case excess risk for ICL-only under representation shift; show minimal rank needed to correct shift; demonstrate tightness (matching rates) up to logs/constant factors.

8. 8. Computational Complexity and Hardness: demonstration subset selection as submodular maximization / NP-hard; approximation guarantees for greedy selection; complexity of adaptation steps; discuss regimes where retrieval dominates cost.

9. 9. Extensions Beyond the Special Case: classification via logistic/softmax linearization; partial labeling (semi-supervised support); federated clients as tasks (adding comm cost); calibration/uncertainty (PAC-Bayes flavored bounds).

10. 10. Experimental Plan (to strengthen the theory): CD-FSL benchmarks, ablations (ICL-only/PEFT-only/hybrid/full finetune), cost metrics, robustness under domain shift and label mapping; failure-case demonstrations aligned with lower bounds; sensitivity to retrieval quality and ordering.

11. 11. Discussion and Future Directions: implications for unified evaluation protocols; guidance for practitioners choosing adaptation under budgets; open questions (beyond linearization, nonparametric shifts, multimodal contexts).

1 Introduction and Motivation

Few-shot learning, as it was commonly operationalized in the episodic meta-learning literature, treated adaptation as a fixed protocol: a learner observes a small support set S_T for a task T, performs a prescribed update (explicit or implicit), and is evaluated on queries Q_T drawn from the same task environment. By 2026, the practical locus of adaptation has shifted. Rather than committing to a single mechanism, practitioners deploy *adaptation menus* built around foundation models: one may (i) keep parameters fixed and rely on in-context learning (ICL) with a carefully chosen set of demonstrations C_T drawn from a retrieval corpus $\mathcal{M}$; (ii) perform parameter-efficient fine-tuning (PEFT) by fitting a constrained update Δ_T (e.g. rank-r adapters) using the few labeled examples in S_T; or (iii) combine the two. Each option has a distinct performance profile and a distinct resource footprint. It is therefore no longer adequate to report accuracy (or loss) alone; one must reason about accuracy jointly with adaptation cost, including context length, retrieval overhead, and training compute.

A second shift is equally decisive: in cross-domain deployments, the main obstacle is not statistical scarcity *per se* but the mismatch between the pre-trained representation and the test domain. When the backbone induces an embedding $\phi : \mathcal{X} \to \mathbb{R}^d$, one should not assume that the same linear head in the ϕ-space is appropriate for all tasks in q_{test}. Instead, we expect a task-specific distortion of the representation, which we model by an unknown operator A_T acting on embeddings. Concretely, we analyze tasks where

$$y \;=\; \langle w_T,\, A_T \phi(x) \rangle \;+\; \xi,$$

with ξ sub-Gaussian and A_T not revealed at test time. The magnitude of shift can be quantified by $\delta := \sup_T \|A_T - I\|_{\text{op}}$. This model is intentionally simple: it isolates the phenomenon that, even if ϕ is strong, the *effective* features at test time may deviate from those for which the backbone is well aligned. In such regimes, additional demonstrations can reduce estimation error but cannot in general eliminate the representation mismatch; conversely, a small amount of parameter updating can explicitly correct the mismatch but introduces nontrivial compute cost and potential instability.

These observations motivate a cost-aware and shift-aware formulation of adaptation. For each task T, an admissible policy chooses a demonstration set $C_T \subseteq \mathcal{M}$ with $|C_T| \le L$ and optionally fits a constrained update Δ_T of rank at most r using at most s steps. We associate a per-task resource cost

$$\text{Cost}(T) \;=\; \alpha L \;+\; \beta r s \;+\; \gamma\, \text{Retr}(L),$$

where α weights context length, β weights PEFT compute, and γ weights system-level overhead such as approximate nearest neighbor lookup time. The central objective is then the cost-regularized risk

$$\mathbb{E}_{T \sim q_{\text{test}}}\big[\mathcal{L}_T(\hat{f}_T)\big] \;+\; \lambda\, \mathbb{E}_{T \sim q_{\text{test}}}\big[\text{Cost}(T)\big],$$

or, equivalently, the constrained-risk problem that traces the error–cost Pareto frontier. The regularization parameter $\lambda > 0$ plays a dual role: it encodes an external budget preference, and it induces a principled tradeoff between performance and adaptation overhead.

Within this framework, the relevant algorithmic question is not "which method wins?" but rather "how should we allocate a limited adaptation budget across heterogeneous tasks?" We address this question by studying a hybrid policy that computes from S_T an observable shift or uncertainty score u_T, retrieves a candidate pool from $\mathcal{M}$, selects a bounded demonstration set C_T, and then *gates* between ICL-only and PEFT (or hybrid) depending on whether the estimated shift is small or large. The gating is designed to approximate the per-task oracle decision that would compare the marginal reduction in expected loss to the marginal increase in cost. Importantly, this view treats retrieval and PEFT as two competing ways to spend adaptation budget: additional context principally decreases estimation error in a fixed feature space, whereas PEFT principally decreases the shift penalty by modifying the effective representation.

Our first contribution is an explicit risk decomposition in a linearized regime that makes this distinction quantitative. When ICL is modeled as ridge regression in the frozen ϕ-space using $K+L$ labeled examples (support plus demonstrations), the excess risk admits an upper bound of the form

$$O\left(\sigma^2 \frac{d}{K+L} \;+\; \delta^2 \|w_T\|^2\right),$$

where the first term is an estimation effect and the second term is an irreducible penalty induced by representation shift. When PEFT is modeled as fitting a rank-r correction to the representation (equivalently a low-rank correction to the effective feature map), the shift penalty becomes compressible and scales as $O(\delta^2 \|w_T\|^2 / r)$, up to estimation terms depending on K. This formalizes a central qualitative fact: increasing L cannot, in general, remove an $O(\delta^2)$ mismatch term if parameters remain frozen, whereas increasing r can.

Our second contribution is a competitive guarantee for cost-aware hybrid gating over a natural restricted policy class Π that includes ICL-only, PEFT-only, and hybrid actions with explicit resource constraints. For each λ, we exhibit a threshold rule (based on statistics computable from S_T) whose achieved cost-regularized objective is within an $O(\log d)$ factor of $\mathrm{OPT}(\lambda)$, the best value attainable by any policy in Π. The role of the factor is to absorb uncertainty in estimating u_T and to account for the discrete choice among actions; the resulting statement is best interpreted as near-Pareto optimality for a broad family of deployment-relevant adaptation strategies.

Our third contribution consists of matching lower bounds that clarify when ICL cannot suffice and when low-rank PEFT is information-theoretically

necessary. In particular, there exist task families for which any ICL-only policy, regardless of the number of demonstrations retrieved, suffers worst-case excess risk $\Omega(\delta^2)$ under frozen ϕ. Moreover, any method restricted to rank-r corrections cannot improve the shift term beyond $\Omega(\delta^2/r)$ in a minimax sense, demonstrating the essential tightness of the rank tradeoff. Finally, we address the computational dimension by showing that optimal demonstration selection is NP-hard in general, while standard greedy selection yields constant-factor approximations when the chosen surrogate utility is monotone submodular. Together, these results justify treating domain shift and cost as first-class objects and motivate hybrid policies that allocate adaptation resources in a task-dependent manner.

2 Related Work and Positioning

The classical meta-learning literature framed few-shot learning as *learning an adaptation algorithm* from a meta-training distribution q_{train} and deploying it on q_{test}, typically under an episodic protocol with support–query splits **????**. In contrast, transfer learning emphasized representation learning by large-scale pretraining followed by task-specific fitting of a lightweight head, with little or no explicit meta-objective **??**. In contemporary foundation-model deployments, this dichotomy is partly dissolved: pretraining provides a strong ϕ, while adaptation is selected from a *menu* that mixes prompting, retrieval, and parameter updates. Our contribution is positioned at this interface, treating the choice among adaptation mechanisms as a decision problem constrained by explicit resource costs.

A second strand concerns the "meta-learning vs. transfer learning" debate in the cross-domain regime. Empirically, methods optimized for within-domain episodes often degrade when q_{test} departs from q_{train}, while strong pretrained backbones can dominate even without sophisticated meta-training **??**. This has prompted a shift from learning fast adaptation dynamics to modeling and correcting representation mismatch **??**. Our formalization makes this focus explicit by introducing a task-dependent shift acting on embeddings; the relevant question becomes not only whether adaptation helps, but which form of adaptation is justified under domain shift and under budget.

Theoretical work on in-context learning (ICL) provides several complementary explanations for why demonstration-conditioned prediction can implement nontrivial learning without parameter updates. One line interprets transformers as approximate Bayesian inference engines over latent task variables, thereby viewing the prompt as conditioning data **??**. A second line establishes connections to implicit gradient descent or to the simulation of iterative algorithms within the forward pass **??**. A third line, most directly aligned with our analysis, shows that in simplified or linearized settings ICL

recovers kernel or (regularized) linear regression in an induced feature space, with the demonstrations serving as training data for an implicit estimator **??**. These results justify modeling ICL as ridge regression in fixed features ϕ, and they also clarify a limitation: if the effective test-time representation differs from the one encoded by ϕ, additional demonstrations can improve estimation but cannot, in general, eliminate misspecification.

Retrieval-augmented prompting and demonstration selection address the fact that, for ICL, *which* examples appear in-context can matter as much as *how many* **??**. Retrieval-augmented generation (RAG) and related memory-augmented methods typically retrieve semantically similar items from a corpus and condition the model on them **??**. For ICL specifically, work has studied nearest-neighbor retrieval in embedding space, diversity-promoting selection, and learned retrievers optimized end-to-end for downstream task loss **??**. From our perspective, retrieval is not a free improvement: it introduces system costs (indexing, latency) and consumes context length, motivating the explicit $\mathrm{Retr}(L)$ and αL terms in the cost model. Moreover, the known hardness of optimal selection motivates surrogate objectives (e.g. information gain or facility-location utilities) for which greedy algorithms provide approximation guarantees **?**.

Parameter-efficient fine-tuning (PEFT) offers a complementary axis of adaptation by modifying the model with a constrained number of trainable parameters. Techniques include adapters, prefix/prompt tuning, and low-rank updates such as LoRA, as well as various gating and sparsity variants **???**. Empirically, PEFT can match or approach full fine-tuning while significantly reducing memory footprint and sometimes improving stability in low-data regimes **??**. Conceptually, PEFT is the natural counterpoint to ICL in our setting: it expends compute (rs updates at rank r for s steps) to modify the effective representation, thereby targeting representation mismatch rather than purely reducing estimation error.

Few-shot PEFT intersects with earlier work on fast adaptation and linear-probe baselines. In vision and language, strong results can be obtained by freezing the backbone and fitting a linear classifier, especially when ϕ is pre-trained at scale **??**. However, under cross-domain shift, linear probing can be brittle, motivating partial adaptation (e.g. tuning only normalization parameters or low-rank components) **??**. Our analysis abstracts this phenomenon by treating PEFT as a constrained correction to the representation; the rank parameter becomes a quantitative knob controlling how much shift can be compensated per unit cost.

Cross-domain few-shot learning (CDFSL) and domain generalization provide additional context for our shift-aware framing. CDFSL benchmarks and methods emphasize transfer from source domains to target domains with limited labeled target data, often combining meta-learning with domain-invariant representation learning or feature normalization **??**. The central obstacle is that the test tasks may deviate in both label semantics and input

distribution, so that performance is governed by the interaction between feature geometry and task-specific decision boundaries. Our model isolates one tractable aspect of this interaction—a task-dependent linear operator acting on embeddings—which permits explicit upper and lower bounds on the residual error induced by shift, and thus a principled comparison between adding context and updating parameters.

We also position the present framework with respect to recent calls for improved evaluation methodology in few-shot and prompting-based systems. Standard reporting of accuracy at a fixed K and a fixed prompting protocol obscures the performance–resource tradeoffs that determine deployability **??**. In particular, retrieval introduces latency and operational complexity; longer context increases inference cost and may reduce throughput; and PEFT introduces training compute and, in some settings, additional storage for per-task adapters. By incorporating these components into a unified objective, we align evaluation with the decision faced by a practitioner: selecting an adaptation strategy subject to budgets.

Finally, the proposed cost-regularized view speaks directly to open problems at the intersection of domain shift and "Green AI" considerations **?**. When performance gains can be purchased by longer context, heavier retrieval, or more gradient steps, it becomes necessary to quantify marginal improvements per unit resource and to identify regimes where additional spend is provably ineffective (e.g. ICL under irreducible shift). Our taxonomy therefore treats adaptation not as a monolithic algorithm but as a portfolio of actions, and it motivates studying Pareto frontiers rather than single operating points. This positioning sets up the formal problem definition in the next section, where we make the objective and admissible policy class explicit.

3 Formal Problem: Cost-Regularized Few-Shot Adaptation (CR-FSA)

We formalize test-time few-shot adaptation as a per-task decision problem in which a deployed foundation model provides a fixed embedding map $\phi : \mathcal{X} \to \mathbb{R}^d$, and the system may additionally spend resources on retrieval, context, and parameter-efficient updates. Our goal is to compare, within a single objective, adaptation mechanisms that (i) consume inference-time context (ICL), (ii) consume training-time compute and parameters (PEFT), and (iii) may be composed (hybrid).

Tasks, episodes, and evaluation. A task T specifies an input–output relationship over $\mathcal{X} \times \mathcal{Y}$ together with a query distribution on (x, y). At

meta-test time we draw $T \sim q_{\text{test}}$ and observe an episodic support set

$$S_T = \{(x_i, y_i)\}_{i=1}^{K} \sim T^K,$$

followed by queries $(x, y) \sim Q_T$ on which we evaluate a loss $\ell(\hat{f}_T(x), y)$. We write the expected query risk of an adapted predictor $\hat{f}_T$ as

$$\mathcal{L}_T(\hat{f}_T) \;:=\; \mathbb{E}_{(x,y)\sim Q_T}\big[\ell(\hat{f}_T(x), y)\big],$$

and the meta-test risk as $\mathbb{E}_{T\sim q_{\text{test}}}[\mathcal{L}_T(\hat{f}_T)]$. We treat S_T as the only labeled information about T; any additional data (e.g. retrieved items) must be obtained without using Q_T labels.

External memory and retrieval. We assume access to a corpus (memory) $\mathcal{M}$, consisting of candidate demonstrations (e.g. labeled examples, instruction–response pairs, or short worked solutions). Given a task T, the system may query an indexed data structure over $\mathcal{M}$ using keys computed from S_T (and, in some deployments, from the query x, but never from its label). The result is a candidate pool $\mathcal{P}_T \subseteq \mathcal{M}$, from which we choose a demonstration set $C_T \subseteq \mathcal{M}$ to include in context. We impose a context-length constraint $|C_T| \leq L$, where L abstracts the token budget as an upper bound on the number of examples that can be inserted into the prompt.

Adaptation actions: ICL, PEFT, and hybrid. For each task T, an adaptation procedure may take one of the following forms.

1. **ICL-only:** select demonstrations C_T with $|C_T| \leq L$ and produce predictions using a fixed backbone (no parameter updates). Operationally, the predictor $\hat{f}_T$ is computed by conditioning on $S_T \cup C_T$ as context; our later analysis will identify a special case in which this corresponds to a regularized linear estimator in the feature space induced by ϕ.

2. **PEFT-only:** choose a parameter-efficient update Δ_T (e.g. adapters, LoRA) from a constrained family and apply it to the base model at test time, fitting Δ_T using S_T for at most s optimization steps. We constrain the update to have rank at most r, reflecting that only $O(dr)$ degrees of freedom are permitted.

3. **Hybrid:** select C_T and fit Δ_T jointly, using both additional context and a parameter update.

In all cases, we emphasize that $\hat{f}_T$ must be measurable with respect to the information available at adaptation time: S_T, the retrieved candidates from $\mathcal{M}$, and any randomness internal to the policy.

Policies. An *adaptation policy* $\pi \in \Pi$ is a (possibly randomized) mapping that, given S_T and oracle access to retrieval over $\mathcal{M}$, outputs an action tuple

$$\pi(S_T) \;=\; \big(C_T, \Delta_T, \hat{f}_T\big),$$

with $|C_T| \le L$ and Δ_T lying in the admissible PEFT family (rank $\le r$, at most s steps). The policy class Π may encode additional restrictions, such as forbidding hybrid composition, fixing L, or limiting the form of the demo-selection rule.

Cost model. We associate to each task T an adaptation cost that captures the dominant controllable resources. We take

$$\mathrm{Cost}(T) \;:=\; \alpha\,|C_T| \;+\; \beta\, r s \;+\; \gamma\,\mathrm{Retr}(|C_T|), \tag{1}$$

where α prices context length (prompt tokens or examples), β prices PEFT optimization effort (rank times steps, standing in for backprop and optimizer overhead), and γ prices retrieval-system overhead. The function $\mathrm{Retr}(L)$ abstracts the scaling of retrieval latency and/or compute with the requested number of items, and may reflect the behavior of an approximate nearest neighbor index or a multi-stage retriever. We allow Retr to be sublinear, linear, or superlinear, depending on system design; the subsequent results will not require a specific functional form beyond monotonicity.

Cost-regularized and constrained objectives. We consider two equivalent ways to pose the design problem. The first is a Lagrangian (regularized) objective: for $\lambda > 0$,

$$\mathrm{Obj}_\lambda(\pi) \;:=\; \mathbb{E}_{T\sim q_{\mathrm{test}}}\big[\mathcal{L}_T(\hat{f}_T^{\pi})\big] \;+\; \lambda\,\mathbb{E}_{T\sim q_{\mathrm{test}}}\big[\mathrm{Cost}^{\pi}(T)\big], \tag{2}$$

and we seek $\inf_{\pi\in\Pi} \mathrm{Obj}_\lambda(\pi)$. Here λ is interpretable as a conversion rate between error and resource expenditure; varying λ traces different operating points.

The second formulation is a constrained risk minimization problem. Given a budget $B \ge 0$, we seek

$$\inf_{\pi\in\Pi}\; \mathbb{E}_{T\sim q_{\mathrm{test}}}\big[\mathcal{L}_T(\hat{f}_T^{\pi})\big] \quad \text{subject to} \quad \mathbb{E}_{T\sim q_{\mathrm{test}}}\big[\mathrm{Cost}^{\pi}(T)\big] \le B. \tag{3}$$

Under standard regularity conditions (e.g. convexity in relaxed policy spaces), (2) is the Lagrangian relaxation of (3); in any case, both objectives provide a principled mechanism to trade off predictive performance against deploy-time resources.

Pareto frontier and the role of hybridization. We define the *error–cost Pareto frontier* induced by Π as the set of achievable pairs

$$\Big(\mathbb{E}_T[\mathrm{Cost}^\pi(T)],\ \mathbb{E}_T[\mathcal{L}_T(\hat{f}_T^\pi)] \Big) \quad \text{for} \quad \pi \in \Pi,$$

restricted to those pairs that are not jointly dominated. In this language, ICL-only, PEFT-only, and hybrid policies correspond to different subsets of Π, and the central question is whether hybrid policies yield strictly better frontiers once domain shift and system costs are accounted for. The remainder of the paper develops a special-case model in which we can (i) upper bound the risk contributions that can be reduced by additional context versus those that require representation correction, (ii) show that optimal or near-optimal policies can be realized by explicit gating rules, and (iii) identify regimes where additional spend is provably futile for certain policy subclasses.

4 A Clean Special-Case Model: The Linearized Foundation-Model Regime

We now isolate a special case in which the interaction between in-context demonstrations and parameter-efficient updates can be studied analytically. The purpose of the model is not to capture the full behavior of a transformer, but to provide a regime in which (i) we can write an explicit query risk, (ii) we can separate estimation effects from representation-shift effects, and (iii) PEFT admits a clean interpretation as a constrained correction to the feature map.

Data model and notation. Fix a pretrained backbone inducing an embedding map $\phi : \mathcal{X} \to \mathbb{R}^d$, and write $z = \phi(x)$. For each task T, we posit a task-specific linear predictor $w_T \in \mathbb{R}^d$ together with an unknown task-specific shift operator $A_T \in \mathbb{R}^{d \times d}$. Labels are generated by

$$y \;=\; \langle w_T, A_T z \rangle \;+\; \xi, \tag{4}$$

where ξ is mean-zero sub-Gaussian noise with parameter σ^2. The Bayes (noise-free) target for task T is therefore the linear function

$$f_T^*(x) \;:=\; \langle w_T, A_T \phi(x) \rangle.$$

We measure shift magnitude by the operator norm deviation from identity,

$$\delta_T \;:=\; \|A_T - I\|_{\mathrm{op}}, \qquad \delta \;:=\; \sup_T \delta_T, \tag{5}$$

and we will be primarily interested in the regime $\delta \ll 1$ (small but non-negligible representation shift).

Feature and covariance assumptions. We assume that, for each task T, the (marginal) embedding distribution has uniformly well-conditioned covariance: if $z \sim P_T$ denotes the embedding of a random task input, then

$$\Sigma_T \;:=\; \mathbb{E}[zz^\top] \quad \text{satisfies} \quad \mu I \preceq \Sigma_T \preceq MI \tag{6}$$

for absolute constants $0 < \mu \le M < \infty$, and z is sub-Gaussian. This is a standard sufficient condition for ridge-regression generalization rates with dimension dependence $\tilde{O}(d/n)$. We emphasize that (6) is imposed on ϕ, not on the unknown shifted features $A_T z$; the latter may be ill-conditioned even if Σ_T is benign when δ is moderate.

Retrieval as additional (possibly imperfect) labeled samples. To focus on the statistical tradeoffs rather than the combinatorics of selecting demonstrations, we introduce an idealized retrieval model. We assume that querying the memory $\mathcal{M}$ with keys derived from S_T returns a candidate pool $\mathcal{P}_T$ containing (at least) m labeled examples distributed as i.i.d. draws from the same task mechanism (4), possibly mixed with irrelevant items. Concretely, one may view $\mathcal{P}_T$ as samples from a mixture $(1-\eta)T + \eta R$, where R is a background distribution and $\eta \in [0, 1)$ quantifies retriever noise. In the cleanest subcase (which we will use when deriving rates), we take $\eta = 0$ so that selecting L demonstrations is equivalent to obtaining L additional labeled samples from T. In later sections we will return to the algorithmic question of selecting $C_T \subseteq \mathcal{P}_T$ under $|C_T| \le L$ when $\eta > 0$ and when selection is computationally constrained.

ICL as ridge regression in fixed features. We specialize to squared loss and linear predictors in the ϕ-feature space. Given any labeled dataset $D = \{(z_i, y_i)\}_{i=1}^n$ (in our setting D will be the concatenation of S_T and C_T, hence $n = K + L$), define the ridge estimator

$$\hat{u}(D) \;:=\; \arg\min_{u \in \mathbb{R}^d} \; \frac{1}{n} \sum_{i=1}^{n} \big(y_i - \langle u, z_i \rangle\big)^2 \;+\; \rho \|u\|_2^2, \tag{7}$$

and the corresponding predictor $\hat{f}_D(x) := \langle \hat{u}(D), \phi(x) \rangle$. In the linearized-foundation-model literature, there are settings in which the in-context computation of a transformer provably implements (or closely approximates) a ridge-type estimator in a fixed feature space; we take (7) as our analytic proxy for ICL-only adaptation.

Under (4), the ridge model $\langle u, z \rangle$ is generally misspecified relative to the shifted target $\langle w_T, A_T z \rangle$ unless the shift can be represented within the fixed feature geometry without altering the embedding map. A convenient way to formalize the resulting irreducible component is to define the best

approximation within the frozen-feature class,

$$u_T^\dagger \;:=\; \arg\min_{u\in\mathbb{R}^d} \; \mathbb{E}\big[(\langle u, z\rangle - \langle w_T, A_T z\rangle)^2\big], \tag{8}$$

so that the excess query risk of any frozen-feature method decomposes into an estimation term (how well we estimate $u_T^\dagger$ from finitely many labels) plus an approximation term (how far $\langle u_T^\dagger, z\rangle$ is from f_T^*). Under (5)–(6), the approximation term can be bounded on the order of $\delta_T^2\|w_T\|_2^2$ (up to conditioning factors), while the estimation term for (7) scales as $\tilde{O}(\sigma^2 d/(K+L))$. This is precisely the regime in which adding demonstrations reduces variance but cannot remove the part of the error attributable to the shift operator.

PEFT as a low-rank correction to the effective feature map. We model PEFT as introducing a task-specific linear correction to the embedding map. Specifically, we take the adapted feature map to be

$$\phi_{\Delta_T}(x) \;:=\; (I + \Delta_T)\phi(x), \tag{9}$$

where $\Delta_T \in \mathbb{R}^{d\times d}$ is constrained to have rank at most r. This abstraction captures, in the linearized regime, the effect of rank-r adapter families (e.g. LoRA-style updates) on the final representation, while keeping the analysis at the level of the embedding space. Given S_T, the PEFT-only estimator fits Δ_T (and, optionally, a linear readout) using at most s gradient steps; in the quadratic model, this corresponds to a regularized least-squares fit over $O(dr)$ free parameters.

The key analytic point is that Δ_T allows us to approximate the unknown shift A_T by a low-rank operator. Writing $A_T = I + E_T$, the best rank-r approximation $E_{T,r}$ (in spectral norm or in the geometry induced by Σ_T) yields a residual $\|E_T - E_{T,r}\|_{\mathrm{op}}^2$ that, for worst-case shifts, scales as δ_T^2/r. Consequently, after fitting a rank-r adapter, the shift-induced component of the excess risk can be reduced from order $\delta_T^2\|w_T\|_2^2$ to order $(\delta_T^2/r)\|w_T\|_2^2$, at the expense of paying estimation error from using only K labeled points and the optimization cost of fitting Δ_T. This is the formal mechanism by which PEFT trades additional compute/parameters for robustness to representation shift.

Taken together, the frozen-feature (ICL) estimator (7) and the adapted-feature (PEFT) model (9) provide a minimal setting in which we can reason about when additional context is statistically useful, when it is provably insufficient, and how rank-constrained correction yields the error–cost tradeoffs that will be exploited by the hybrid algorithms in the next section.

5 Algorithms: Hybrid Adaptation via Cost-Aware Gating

We now describe (i) an analytic hybrid policy in the clean special case of Section 4, and (ii) a practical policy template that instantiates the same decision logic under realistic retrieval noise and strict per-task budgets. The common structure is: from the support set we compute a cheap *shift/uncertainty score* u_T, we use retrieval to assemble a small pool of candidate demonstrations, we select C_T under $|C_T| \le L$, and we decide whether to incur the additional cost of fitting a rank-r update Δ_T.

5.1 An analytic hybrid policy in the clean retrieval subcase

We first consider the idealization in which retrieval returns i.i.d. samples from the same task mechanism, so that choosing $|C_T| = L$ is equivalent to observing L additional labeled samples. In this regime, the upper bounds of Thm 1 directly induce a per-task *proxy objective* for each action. Writing $\mathrm{Cost}(T) = \alpha L + \beta rs + \gamma\,\mathrm{Retr}(L)$, we define the following bound-driven surrogates:

$$\mathrm{UB}_T^{\mathrm{ICL}}(L) := c_1\,\sigma^2 \frac{d}{K+L} \;+\; c_2\,u_T, \tag{10}$$

$$\mathrm{UB}_T^{\mathrm{PEFT}}(r) := c_1\,\sigma^2 \frac{d}{K} \;+\; c_3\,\frac{u_T}{r}, \tag{11}$$

$$\mathrm{UB}_T^{\mathrm{hyb}}(L,r) := c_1\,\sigma^2 \frac{d}{K+L} \;+\; c_3\,\frac{u_T}{r}, \tag{12}$$

for absolute constants $c_i > 0$. Here u_T is any statistic computed from S_T that upper bounds (up to constants and conditioning factors) the shift-induced term $\delta_T^2 \|w_T\|_2^2$; we will provide concrete estimators below. Given $\lambda > 0$, the analytic policy chooses the action minimizing the corresponding cost-regularized proxy:

$$(L_T, r_T, \mathrm{mode}_T) \in \arg \min_{\substack{0 \le L \le L_{\max} \\ 0 \le r \le r_{\max}}} \Big\{ \min\big(\mathrm{UB}_T^{\mathrm{ICL}}(L) + \lambda(\alpha L + \gamma\,\mathrm{Retr}(L)),\; \mathrm{UB}_T^{\mathrm{PEFT}}(r) + \lambda(\beta rs),\; \mathrm{UB}_T^{\mathrm{hyb}}(\tag{13}$$

In the clean case where $\mathrm{Retr}(L)$ is nondecreasing and L, r are small enough to allow enumeration, (13) is computationally trivial. More importantly, it makes explicit that the *gating* is determined by the single scalar u_T: for fixed $(K, d, \sigma^2, \lambda)$ and fixed system costs (α, β, γ), the preference ordering among ICL, PEFT, and hybrid changes at thresholds proportional to u_T. For example, comparing (10) and (11) yields a sufficient condition of the

form

$$u_T \gtrsim \frac{\sigma^2 d}{\frac{1}{r} - 0} + \lambda\Big(\beta rs - \alpha L - \gamma\,\mathrm{Retr}(L)\Big) \qquad \Longrightarrow \qquad \text{prefer PEFT to ICL,} \tag{14}$$

with constants suppressed. In words: if the inferred shift is large enough that the u_T/r correction dominates the additional estimation term, then PEFT is worth paying for; if u_T is small, additional context is preferred since it reduces only the estimation component. The hybrid mode becomes favorable when K is small (so $\sigma^2 d/K$ is large) and u_T is non-negligible (so some correction is needed), exactly matching the qualitative behavior of Thm 1.

5.2 A practical gating policy template

We now give a deployable template that approximates (13) while respecting the "no leakage" constraint (adaptation uses only S_T and corpus items independent of query labels).

Step 1: compute a shift/uncertainty score u_T. We require u_T to be cheap and monotone in representation mismatch. Three concrete choices are:

1. *Residual-based score.* Fit the frozen-feature ridge estimator on S_T, obtaining $\hat{u}_T$. Define

$$u_T \;:=\; \frac{1}{K} \sum_{(x_i, y_i) \in S_T} \big(y_i - \langle \hat{u}_T, \phi(x_i) \rangle\big)^2 \;-\; \widehat{\sigma}^2,$$

   clipped at 0. In the well-specified case this concentrates near 0; under shift it captures systematic error beyond noise.

2. *Geometry-based score.* Maintain corpus-level source statistics $(\hat{\mu}_0, \hat{\Sigma}_0)$ for embeddings. Let $\bar{z}_T$ be the mean support embedding and let $\hat{\Sigma}_T$ be the empirical covariance on S_T. Define

$$u_T \;:=\; \|\hat{\Sigma}_0^{-1/2}(\bar{z}_T - \hat{\mu}_0)\|_2^2 \;+\; \|\hat{\Sigma}_0^{-1/2}(\hat{\Sigma}_T - \hat{\Sigma}_0)\hat{\Sigma}_0^{-1/2}\|_{\mathrm{F}}^2.$$

   This detects domain shift even when labels are scarce or noisy.

3. *Influence/condition score.* Use the ridge design matrix $Z \in \mathbb{R}^{K \times d}$ with rows $z_i^\top$. Set $u_T := \mathrm{tr}\big((Z^\top Z + K\rho I)^{-1}\big)$, which proxies predictive variance; large values indicate the frozen representation is poorly aligned with the task-relevant directions.

Any of these can be computed in one pass over S_T (and possibly with precomputed corpus statistics).

Step 2: retrieve a candidate pool and select demonstrations. Given a pool $\mathcal{P}_T = \mathrm{Retrieve}(\mathcal{M}, S_T, m)$, we choose $C_T \subseteq \mathcal{P}_T$ by maximizing a monotone submodular surrogate under $|C_T| \le L$. A canonical choice consistent with ridge-risk reduction is log-determinant information gain:

$$U(C) \;:=\; \log\det\Big(I + \tfrac{1}{\rho} \sum_{(x,y)\in S_T \cup C} \phi(x)\phi(x)^\top\Big), \tag{15}$$

for which greedy selection attains a $(1 - 1/e)$-approximation when submodularity conditions hold. In practice, we evaluate marginal gains using incremental Cholesky updates in a low-dimensional sketch (e.g. random projections of $\phi(x)$) to keep selection cost sublinear in d.

Step 3: decide ICL vs. PEFT vs. hybrid under a budget. We implement a two-threshold gate. First, we choose L_T by trading off $\alpha L + \gamma\,\mathrm{Retr}(L)$ against the marginal reduction in an uncertainty proxy (e.g. the drop in $\mathrm{tr}((Z^\top Z + \rho I)^{-1})$ after adding candidates). Second, we decide whether to train an adapter by comparing u_T to a learned or calibrated threshold $\tau(\lambda, B)$ derived from (14). Concretely:

$$u_T \le \tau \;\Rightarrow\; \Delta_T = 0 \text{ (ICL-only)}, \qquad u_T > \tau \;\Rightarrow\; \text{fit rank-}r \text{ adapter for } s \text{ steps}.$$

When a strict per-task budget B is imposed, we choose the largest feasible L_T and (r_T, s_T) satisfying $\alpha L_T + \beta r_T s_T + \gamma\,\mathrm{Retr}(L_T) \le B$, and we apply the gate within that feasible set; equivalently, we run the Lagrangian form with λ tuned so that the realized average cost matches B.

This template mirrors the analytic policy while isolating all implementation dependence into: (i) the design of u_T, (ii) the retrieval and selection heuristic for C_T, and (iii) the calibration of τ and the mapping from budget to (L, r, s). In Section 6 we will upper bound the resulting excess risk by combining (a) estimation terms controlled by K and L and (b) shift terms controlled by u_T and the chosen rank r, and we will relate the gate to near-Pareto optimality within the restricted policy class.

6 Upper Bounds: Generalization and Cost–Error Tradeoffs

We now turn the bound-driven surrogates of Section 5 into explicit excess-risk guarantees and an error–cost characterization. For a task T, we write the excess query risk of an adapted predictor $\hat{f}_T$ as

$$\mathcal{E}_T(\hat{f}_T) \;:=\; \mathbb{E}\big[\mathcal{L}_T(\hat{f}_T)\big] - \mathcal{L}_T(f_T^*),$$

where the expectation is over the support/query sampling and the noise ξ, and f_T^* denotes the Bayes-optimal predictor within the assumed linearized model.

A three-term decomposition. In the linearized regime $y = \langle w_T, A_T\phi(x)\rangle + \xi$, the dominant contributions to $\mathcal{E}_T$ separate into:

1. *Estimation* due to finite labeled data used by the final linear predictor in feature space;
2. *Shift* due to using the unshifted representation ϕ when the task is realized through $A_T\phi$;
3. *Approximation/optimization* due to constraining the adaptation mechanism (e.g. rank-r adapters and s optimization steps).

Concretely, if $\hat{f}_T^{\mathrm{ICL}}$ is ridge regression in frozen features ϕ using $K+L$ labeled examples (support plus demonstrations), standard sub-Gaussian generalization bounds yield

$$\mathcal{E}_T(\hat{f}_T^{\mathrm{ICL}}) \;\le\; O\!\left(\sigma^2\,\frac{d}{K+L}\right) \;+\; O\!\left(\|(A_T - I)^\top w_T\|_2^2\right), \tag{16}$$

where the first term is the familiar d/n rate (up to logs and covariance conditioning), and the second term is a shift penalty which in particular is bounded by $O(\delta^2\|w_T\|_2^2)$ under $\|A_T - I\|_{\mathrm{op}} \le \delta$. This makes precise the qualitative limitation: increasing L suppresses only estimation, not shift.

If instead we fit a rank-r adapter, we may view it as learning a rank-r correction to the effective feature map. In the idealized convex setting (linear ridge over adapted features), we obtain

$$\mathcal{E}_T(\hat{f}_T^{\mathrm{PEFT}}) \;\le\; O\!\left(\sigma^2\,\frac{d}{K}\right) \;+\; O(\mathrm{Tail}_r(T)) \;+\; \mathrm{OptErr}(s), \tag{17}$$

where $\mathrm{Tail}_r(T)$ measures how well the shift direction can be corrected by a rank-r update. Under the isotropic truncation model used in Thm 1, $\mathrm{Tail}_r(T) = O(\delta^2\|w_T\|_2^2/r)$. More generally, if $(\sigma_j)_{j\ge1}$ are singular values of $A_T - I$, then one may take $\mathrm{Tail}_r(T)$ proportional (up to covariance factors) to $\sum_{j>r}\sigma_j^2\,\|w_T\|_2^2$, emphasizing that low-rank PEFT is effective precisely when the shift is concentrated on a low-dimensional subspace. Finally, $\mathrm{OptErr}(s)$ captures imperfect fitting with s steps; for strongly convex objectives and appropriate step size, one may take $\mathrm{OptErr}(s) \le (1-\eta\mu)^s \cdot \mathrm{OptErr}(0)$, while for stochastic optimization it is standard to obtain $\mathrm{OptErr}(s) = O(1/s)$ up to variance terms.

The hybrid predictor $\hat{f}_T^{\mathrm{hyb}}$, which uses both demonstrations and an adapter, inherits the improved estimation term $O(\sigma^2 d/(K+L))$ and the improved shift term $\mathrm{Tail}_r(T)$, plus the same optimization residual:

$$\mathcal{E}_T(\hat{f}_T^{\mathrm{hyb}}) \;\le\; O\!\left(\sigma^2\,\frac{d}{K+L}\right) \;+\; O(\mathrm{Tail}_r(T)) \;+\; \mathrm{OptErr}(s). \tag{18}$$

Cost–error optimization and parameter dependence. We now connect (16)–(18) to the cost model

$$\mathrm{Cost}(T) \;=\; \alpha L + \beta r s + \gamma\,\mathrm{Retr}(L).$$

For a fixed task T, if we temporarily ignore $\mathrm{Retr}(L)$ and discretization effects, the minimizers of the cost-regularized surrogates exhibit explicit scaling. For ICL-only, minimizing $c\,\sigma^2 d/(K+L) + \lambda\alpha L$ yields

$$L_T^* \;\approx\; \Big(\sqrt{\tfrac{c\,\sigma^2 d}{\lambda\alpha}} - K\Big)_+,$$

illustrating that (i) larger λ forces shorter context, (ii) larger d or σ^2 makes additional demonstrations more valuable, and (iii) when K is already large, the optimal incremental context quickly drops to zero. For PEFT-only with $\mathrm{Tail}_r(T) \asymp u_T/r$ (as in Thm 1), minimizing $c' u_T/r + \lambda\beta r s$ yields

$$r_T^* \;\approx\; \sqrt{\tfrac{c' u_T}{\lambda\beta s}},$$

so that large inferred shift u_T justifies higher rank, whereas expensive adaptation (large β or s) pushes toward smaller rank. Hybrid choices combine these scalings, with L responding primarily to the estimation term and r responding primarily to the shift term. The dependence on δ enters through u_T: in the worst case $u_T \asymp \delta^2\|w_T\|_2^2$, and hence the best achievable shift penalty under rank-r correction scales as δ^2/r up to constants.

When retrieval is imperfect, we may add an explicit *retrieval bias* term $\varepsilon_{\mathrm{retr}}(m, L)$ to (16) and (18) capturing the discrepancy between the distribution of retrieved items and the true task distribution; in the clean subcase $\varepsilon_{\mathrm{retr}} = 0$. This term increases with aggressive selection from a small or mismatched corpus and is the point at which system design (indexing, filtering, diversity constraints) enters the statistical guarantee.

Near-Pareto optimality among restricted policy classes. We finally relate the gating policy to the cost–error Pareto frontier. Consider the restricted policy class Π of per-task choices among ICL-only, PEFT-only, and hybrid, each with bounded (L, r, s) as in Section 5. For $\lambda > 0$, define the cost-regularized value of a policy $\pi \in \Pi$ by

$$V_\lambda(\pi) \;:=\; \mathbb{E}_{T\sim q_{\mathrm{test}}}\big[\mathcal{L}_T(\hat{f}_T^{\pi})\big] + \lambda\,\mathbb{E}_{T\sim q_{\mathrm{test}}}\big[\mathrm{Cost}^{\pi}(T)\big].$$

Thm 2 asserts that an explicit threshold rule using observable statistics from S_T achieves

$$V_\lambda(\pi_{\mathrm{gate}}) \;\le\; O(\log d)\cdot \inf_{\pi\in\Pi} V_\lambda(\pi).$$

The proof compares (task by task) the upper bounds (16)–(18) plus costs, and uses concentration to show that the estimated shift score u_T is sufficiently accurate to select the same mode as the oracle minimizer up to logarithmic slack. By standard Lagrangian duality, varying λ traces a Pareto

frontier between expected loss and expected cost; thus, within Π, the gating policy is near-Pareto optimal up to the same $O(\log d)$ factor in the regularized objective, and hence attains (up to logarithmic degradation) the best achievable error at any prescribed average cost level.

In summary, the upper bounds separate the roles of K, L, r, s, and δ: context primarily combats estimation through $K + L$, adapters primarily combat shift through rank r (and are limited by optimization budget s), and the gate selects the least costly mechanism whose bound is competitive for the observed task statistics.

7 Lower Bounds and Impossibility

The upper bounds of Section 6 are only meaningful insofar as they are unavoidable: we now show that the dependence on the shift magnitude δ cannot, in general, be eliminated by *frozen-representation* adaptation, and that the $1/r$ improvement afforded by rank-r correction is essentially optimal (up to logarithmic factors and conditioning constants). Throughout we emphasize *minimax* and *worst-case* statements, since the goal is to delineate what cannot be achieved uniformly over a class of cross-domain tasks.

ICL-only cannot uniformly eliminate representation shift. We formalize "ICL-only" as any policy that, upon observing (S_T, C_T), outputs a predictor whose dependence on inputs x is mediated solely through the fixed embedding $\phi(x)$; in particular, the policy may be arbitrarily powerful computationally and may choose demonstrations C_T adaptively, but it performs no parameter update that changes the effective representation seen at test time. Thm 3 states that this restriction alone implies an irreducible shift floor: there exists a task family with $\|A_T - I\|_{\mathrm{op}} = \delta$ for which every such policy suffers excess risk $\Omega(\delta^2)$ in the worst case.

A convenient way to see why such a phenomenon occurs is to construct two hypotheses T_0, T_1 whose induced distributions over the *observable* random variables (S_T, C_T) are (nearly) indistinguishable, but whose Bayes predictors differ on the query distribution by a fixed amount. Concretely, we choose a feature distribution for $\phi(x)$ supported on a low-dimensional subspace $U \subset \mathbb{R}^d$ on which A_{T_0} and A_{T_1} act identically, while ensuring that the query distribution has nontrivial mass on another subspace $U^\perp$ where the two shift operators differ. Since an ICL-only predictor is a measurable function of (S_T, C_T) and $\phi(\cdot)$, it cannot infer which of the two tasks it is facing (the observations are the same), and hence must incur error on one of them. In the simplest Gaussian instantiation, one may take $\phi(x) \sim \mathcal{N}(0, I_d)$ for queries but $\phi(x) \sim \mathcal{N}(0, P_U)$ for the labeled information available to the policy (support plus any admissible retrieved demonstrations), and then

choose

$$A_{T_0} = I, \qquad A_{T_1} = I + \delta\, uv^\top,$$

with $u \in U^\perp$, $v \in U^\perp$, $\|u\|_2 = \|v\|_2 = 1$, and w_T aligned with u. The two tasks agree on all labeled observations (since those live in U), yet differ on the query labels through the component $v^\top \phi(x)$ which is invisible during adaptation. A standard Le Cam two-point argument then yields

$$\inf_{\text{ICL-only } \pi} \sup_{T \in \{T_0, T_1\}} \mathcal{E}_T(\hat{f}_T^\pi) \gtrsim \delta^2,$$

where the constant depends only on the signal/noise normalization. This is the content of Thm 3: even with arbitrarily many demonstrations ($L \to \infty$) drawn from an arbitrarily large corpus, if the policy cannot change the representation used at test time, there exist shifts that are statistically undetectable from the adaptation information and hence cannot be corrected.

Retrieval does not circumvent the lower bound without representation correction. The same indistinguishability perspective explains why optimizing C_T cannot, by itself, eliminate the worst-case shift penalty. If the retrieval mechanism is constrained to return demonstrations whose embeddings lie in the same "observable" sigma-algebra (for example, governed by the same corpus distributional support), then there exist shifts that are *orthogonal* to every piece of information the policy can condition on. In such cases, increasing L improves estimation in the accessible subspace but leaves a bias term whose magnitude is controlled by $\|A_T - I\|_{\text{op}}$, yielding precisely the kind of δ-dependent floor highlighted in Section 6. This clarifies the logical role of PEFT in our framework: it is not merely a data-efficiency device, but a mechanism for changing the effective hypothesis class so that previously unidentifiable shift directions become representable.

A rank requirement: no method with rank-r correction can beat δ^2/r in general. We next justify the rate in the PEFT shift term by showing that low-rank correction has an inherent approximation barrier. Thm 4 asserts that for a suitable family of shifts, any method constrained to a rank-r correction incurs minimax excess risk $\Omega(\delta^2/r)$. One may view this as a spectral "packing" statement: if the unknown shift $A_T - I$ spreads its energy across $\Theta(r)$ nearly orthogonal directions, then any rank-r update must leave a nontrivial tail, and the resulting uncorrected component produces error proportional to that tail energy.

A concrete construction is to take $A_T - I$ to be a random sign combination of m orthogonal rank-one directions,

$$A_T - I \;=\; \frac{\delta}{\sqrt{m}} \sum_{j=1}^{m} \varepsilon_j\, u_j v_j^\top, \qquad \varepsilon_j \in \{\pm 1\},$$

with $\{u_j\}_{j=1}^m$ and $\{v_j\}_{j=1}^m$ orthonormal families, and choose $m \asymp r$ (or larger). Any rank-r approximation can capture at most r directions, leaving residual operator norm and Frobenius mass on the remaining components. When w_T has nontrivial projection on the affected subspace, the induced error behaves like the squared magnitude of the uncorrected shift component, which yields a lower bound of order δ^2/r after averaging over the random signs (or taking a worst-case realization). Technically, one couples this approximation argument with a minimax lower bound for estimating a low-rank matrix from K samples, producing the stated dependence on δ and r up to universal constants.

Tightness and interpretation. Taken together, Thm 3 and Thm 4 show that the qualitative behavior in Thm 1 is essentially sharp: frozen-feature adaptation cannot remove worst-case shift ($\Omega(\delta^2)$), whereas allowing rank-r representation correction yields the best possible uniform improvement factor r in the shift-induced term ($\Theta(\delta^2/r)$). In particular, the upper bound $\mathcal{E}_T(\hat{f}_T^{\mathrm{PEFT}}) \lesssim \sigma^2 d/K + \delta^2 \|w_T\|_2^2/r$ cannot be improved in its (δ, r) dependence without strengthening assumptions on the task family (e.g., stronger structure on A_T, alignment between w_T and the top singular directions of $A_T - I$, or access to additional supervision that reveals the hidden shift directions).

Finally, these lower bounds justify the design of a *cost-aware* hybrid gate: when the estimated shift is small, PEFT is unnecessary because the δ-dependent term is dominated by estimation; when the shift is large, ICL-only is information-theoretically insufficient in the worst case, so spending budget on rank (and steps) is the only route to uniformly improved error. Having established what is statistically achievable, we next turn to what is computationally achievable: even when additional demonstrations would help, selecting them optimally is, in general, intractable, motivating the approximation-oriented retrieval and selection procedures analyzed in the next section.

8 Computational Complexity and Hardness

We now account for the computational resources required by hybrid adaptation and clarify why, even when additional demonstrations are statistically beneficial, selecting them *optimally* is generally intractable. This complements the statistical lower bounds of Section 7 by separating what is information-theoretically impossible from what is computationally prohibitive.

Per-task resource model. Recall that an admissible policy, upon observing the support set S_T of size K, may (i) embed inputs via ϕ, (ii) query a

retrieval structure over $\mathcal{M}$ to obtain a candidate pool, (iii) select a context set C_T with $|C_T| \leq L$, and (iv) optionally fit a rank-r adapter Δ_T for at most s steps. In our RAM-style abstraction, the dominant costs decompose into forward embedding evaluations, retrieval and selection overheads, and (if enabled) adapter optimization. Writing T_ϕ for the time of one backbone forward pass, the embedding stage costs $O((K+L)T_\phi)$ once we have committed to a context of size L; if embeddings for $\mathcal{M}$ are precomputed, then retrieval need only operate in $\mathbb{R}^d$.

A typical retrieval pipeline first returns a pool $\mathcal{P}_T$ of size m (often $m \gg L$) using an approximate nearest-neighbor (ANN) index. We keep this cost abstract as $\mathrm{Retr}(m)$, since it depends on the data structure (HNSW, IVF-PQ, LSH) and on the desired recall. Once a pool is obtained, selecting L elements from $\mathcal{P}_T$ can range from trivial (top-L by similarity) to expensive (solving a combinatorial optimization).

Finally, if we run PEFT, the update Δ_T has $O(dr)$ degrees of freedom (e.g., LoRA-style low-rank factors) and is optimized for s steps using only S_T. Ignoring constant factors from backpropagation, a crude but informative bound is $O(sdr)$ arithmetic operations for adapter-only updates; the full wall-clock cost depends on the implementation (activation checkpointing, batch size K, and whether multiple layers are adapted), but the scaling in (s, d, r) captures the essential tradeoff.

Hardness of optimal demonstration selection. Even in the linearized regime where ICL corresponds to ridge regression in fixed features, the problem

$$\min_{C \subseteq \mathcal{P}_T:\ |C| \leq L} \ \mathbb{E}\big[\mathcal{L}_T(\hat{f}_T^{\mathrm{ICL}}(S_T \cup C))\big]$$

is computationally difficult in general. Thm 5 formalizes this by showing NP-hardness of selecting a size-L subset C that optimizes downstream risk (or equivalently maximizes a natural notion of improvement). The reduction is standard in spirit: one encodes a Max-Cover or Facility-Location instance into a set of candidate demonstrations, and defines a task-dependent utility such that including a demonstration corresponds to "covering" an element (or opening a facility) in the underlying instance. Any algorithm that solved the demonstration selection problem exactly in polynomial time would therefore solve an NP-hard combinatorial problem, implying $P = NP$.

We emphasize that this hardness is not an artifact of exotic loss functions or nonconvex training. It persists even when the downstream predictor is a closed-form ridge solution in fixed features: the combinatorial difficulty arises from the interaction between *subset choice* and the *geometry* of the resulting design matrix in $\mathbb{R}^d$.

Submodular surrogates and greedy approximation. Given NP-hardness, we must either restrict the instance family or adopt approximation algo-

rithms. A common and analytically convenient approach is to replace the true risk objective by a surrogate utility that is (approximately) monotone submodular. Concretely, let $Z(C) \in \mathbb{R}^{|C| \times d}$ be the matrix whose rows are embedded features $\phi(x)$ of candidates in C, and consider the information-gain objective

$$U(C) \;:=\; \log\det\Big(I + \tfrac{1}{\eta} Z(C)^\top Z(C)\Big),$$

for some $\eta > 0$. Under mild conditions, U is normalized, monotone, and submodular as a set function of C. Intuitively, the marginal gain of adding a new demonstration decreases as the selected set already spans the relevant directions in feature space. When such a surrogate is used (possibly with task-dependent reweighting based on S_T), the classic Nemhauser–Wolsey theorem yields that greedy selection achieves a $(1 - 1/e)$-approximation to the optimal $U(C)$ subject to $|C| \le L$. This is precisely the approximation guarantee cited in Thm 5.

From an implementation perspective, greedy selection requires evaluating L rounds of marginal gains over a pool of size m, leading to $O(Lm)$ utility evaluations. When U admits incremental updates (e.g., via the matrix determinant lemma and maintaining a Cholesky factor of $I + \frac{1}{\eta} Z^\top Z$), each marginal gain can be computed in $O(d^2)$ time naively or faster with low-rank structure and cached inverses; in practice one often operates in a reduced dimension or uses diagonal/structured approximations to keep the per-evaluation cost small. Abstractly, we summarize this as $O(Lm\, c_u)$, where c_u is the per-candidate marginal-gain cost.

Overall per-task complexity and dominant regimes. Putting the pieces together, a representative per-task time bound is

$$O\Big((K + L)T_\phi \;+\; \mathrm{Retr}(m) \;+\; Lm\, c_u \;+\; sdr\Big),$$

where the last term is incurred only if the gate triggers PEFT. This decomposition highlights the regimes in which different components dominate:

- *Large corpus / high recall: retrieval dominates.* When $|\mathcal{M}|$ is large and high recall is demanded, $\mathrm{Retr}(m)$ and subsequent selection over a large pool m can dominate, even if r and s are modest. In such cases, reducing m, using cheaper similarity keys derived from S_T, or replacing greedy with lighter heuristics can provide substantial savings with limited accuracy loss.

- *Large drs: adapter optimization dominates.* When we allow higher-rank adapters or multiple steps, the sdr scaling can exceed retrieval costs, particularly when the backbone is large and the adapter back-propagation is not fully optimized. This is the regime in which the cost term βrs is an accurate proxy for wall-clock time, and the gate meaningfully trades compute for reduced shift penalty (cf. Thm 1).

- *Small m and small r: embedding evaluation dominates.* If retrieval is lightweight (small m) and PEFT is either off or very small, then the repeated forward passes $(K+L)T_\phi$ may be the bottleneck, motivating embedding caching and careful batching.

Thus, the cost-aware design is not merely a modeling convenience: the same parameters (L, r, s) that control statistical error also control computational feasibility, and the hybrid gate can be interpreted as a mechanism for allocating time across retrieval, selection, and optimization in a task-adaptive manner.

9 Extensions Beyond the Special Case

Our analysis above is stated in a squared-loss linear model on fixed embeddings, with PEFT modeled as a low-rank correction that reduces representation shift. We now indicate several extensions in which the same error–cost tradeoffs and gating logic persist, albeit with modified technical tools. Throughout, we maintain the episodic protocol: the policy observes only S_T (and a retrieval corpus $\mathcal{M}$), produces an adapted predictor $\hat{f}_T$, and is evaluated on Q_T.

Classification via logistic/softmax linearization. For binary classification, suppose that conditional on x the label satisfies a logistic model

$$\mathbb{P}(y = 1 \mid x, T) \;=\; \sigma\big(\langle w_T, A_T\phi(x)\rangle\big), \qquad \sigma(t) = \tfrac{1}{1+e^{-t}},$$

and we evaluate the expected logistic loss $\ell(y,t) = \log(1+\exp(-yt))$ with $y \in \{\pm 1\}$. In the regime where the logits remain in a bounded range on the support and query distributions (a standard condition ensuring strong convexity of the empirical risk), ridge logistic regression in the fixed embedding behaves similarly to ridge least squares, with the substitution of the feature covariance by the Hessian-weighted covariance. Concretely, letting H_T denote the Hessian of the population logistic risk at the task optimum, we may treat $\sqrt{H_T}\, A_T\phi(x)$ as the effective feature map; estimation error scales as $O(\frac{d}{K+L})$ up to condition numbers, while the shift penalty depends on $\|A_T - I\|_{\mathrm{op}}$ through perturbation bounds for H_T.

For multi-class classification with N classes, we similarly consider a softmax model with class weight vectors $w_{T,1}, \ldots, w_{T,N} \in \mathbb{R}^d$ and logits $\langle w_{T,c}, A_T\phi(x)\rangle$. One convenient reduction is to linearize the softmax cross-entropy around a reference point w^0 (e.g., the pretrained head or a meta-learned initialization) and analyze the local quadratic approximation given by the Fisher information. This yields, on a per-task basis, a generalized ridge problem in a feature space of dimension $d(N-1)$ with a block structure. In this view, ICL corresponds to fitting a linearized classifier in frozen

features using $K + L$ examples, while PEFT again acts by modifying the effective features through a low-rank correction. The resulting bounds inherit the same qualitative structure as Thm. 1: an estimation term controlled by the number of labeled examples used by the in-context procedure, and a shift term that PEFT reduces at a rate governed by the rank r, now with constants depending on softmax curvature and class separability.

Partially labeled and semi-supervised support sets. In many few-shot episodes, only a subset of support points carry labels. Let S_T^ℓ denote K_ℓ labeled examples and S_T^u denote K_u unlabeled examples, with $K_\ell + K_u = K$. In the linearized regression regime, unlabeled points can improve adaptation in two distinct ways. First, they provide a better estimate of the task-specific feature covariance under the shifted representation $A_T\phi(x)$. If we perform ridge regression with a covariance preconditioner estimated from S_T^u (or use a whitening transform derived from unlabeled embeddings), then the sensitivity of the predictor to shift directions can be reduced even before invoking PEFT. Second, unlabeled points can be used transductively: one may choose C_T and/or Δ_T to minimize a supervised objective on S_T^ℓ plus an unsupervised regularizer on S_T^u (e.g., entropy minimization or consistency under small perturbations in embedding space). In the linear model, a natural regularizer is $\sum_{x \in S_T^u} \|\hat{w}^\top \phi(x)\|^2$, which biases toward predictors stable on the task marginal.

From the standpoint of our hybrid policy, unlabeled data also sharpens the gating statistic u_T. A shift score based only on K_ℓ labeled points can be noisy; incorporating S_T^u allows us to estimate dispersion and out-of-domain distances in $\mathbb{R}^d$ with variance shrinking like $1/K$. This reduces the probability of choosing the wrong action (ICL-only versus PEFT-enabled) and hence improves the competitive ratio analysis implicit in Thm. 2. Importantly, using S_T^u does not alter the cost accounting except through additional embedding computations (already present in our model), so the same cost regularizer $\alpha L + \beta r s + \gamma\,\mathrm{Retr}(L)$ applies.

Federated clients as tasks and communication cost. A practically salient instantiation of q_{test} is a federated population, where each task T corresponds to a client domain with its own support set S_T (often non-IID across clients). In this setting, adaptation can be executed locally, while retrieval and model parameters may be hosted centrally. The hybrid framework extends by augmenting the per-task cost with a communication term. For example, if the policy transmits an adapter update Δ_T (or sufficient statistics derived from S_T) to a server, we may define

$$\mathrm{Cost}_{\text{fed}}(T) \;:=\; \alpha L + \beta r s + \gamma\,\mathrm{Retr}(L) + \rho\,\mathrm{Comm}(T),$$

where $\mathrm{Comm}(T)$ measures bits sent/received and ρ converts bandwidth into a cost weight. Low-rank adapters are naturally communication-efficient:

sending LoRA factors requires $O(dr)$ scalars rather than $O(d^2)$, and in many deployments one can further quantize or sparsify these factors. Retrieval can be handled either centrally (client sends a query key derived from S_T) or locally (client maintains a private subset of $\mathcal{M}$); the former introduces privacy and communication constraints, while the latter introduces memory constraints. In either case, the gating logic remains: we compare the predicted marginal loss reduction from enabling PEFT (or enlarging L) against the marginal increase in $\mathrm{Cost}_{\mathrm{fed}}(T)$, now including $\mathrm{Comm}(T)$.

Calibration, uncertainty, and PAC-Bayes-style bounds. Finally, the hybrid policy is only as reliable as its ability to assess uncertainty and shift. Beyond point estimates of u_T, we can maintain a distribution over task predictors to obtain calibrated probabilities and principled decision rules. One route is PAC-Bayes: for each task T, we consider a posterior $\mathcal{Q}_T$ over predictors (e.g., over w_T in the frozen-feature model, or over adapter parameters in the PEFT model) obtained from S_T, with a prior $\mathcal{P}$ derived from meta-training. Standard PAC-Bayes inequalities then yield, with high probability over S_T,

$$\mathbb{E}_{f\sim\mathcal{Q}_T}\big[\mathcal{L}_T(f)\big] \;\le\; \widehat{\mathcal{L}}_{S_T}(\mathcal{Q}_T) \;+\; O\!\left(\sqrt{\tfrac{\mathrm{KL}(\mathcal{Q}_T\|\mathcal{P})+\log(1/\delta_0)}{K}}\right),$$

for a confidence parameter δ_0. In our context, $\mathrm{KL}(\mathcal{Q}_T\|\mathcal{P})$ acts as a complexity penalty that is sensitive to whether we (i) rely on ICL-only (small posterior shift), (ii) fit a larger-rank adapter (larger parameter movement), or (iii) combine both. Thus the same ingredients that enter $\mathrm{Cost}(T)$ also enter an uncertainty-aware upper bound on risk, providing a coherent basis for gating: we can choose the action minimizing an explicit upper confidence bound plus $\lambda\,\mathrm{Cost}(T)$. Separately, calibrated predictive uncertainty can be used to order or filter demonstrations (preferring those that reduce posterior variance in directions most relevant to S_T), thereby connecting selection heuristics to provable generalization control.

These extensions suggest that the linearized squared-loss case is best viewed as a minimal template: once we can express adaptation as controlling an estimation term and a shift/approximation term under resource constraints, the same hybrid mechanism and error–cost frontier persist across classification, semi-supervision, federated deployment, and uncertainty-aware decision-making.

10 Experimental Plan: Empirically Stress-Testing the Theory

We outline an experimental program whose purpose is to (i) instantiate the objective $\mathbb{E}[\mathcal{L}_T(\hat{f}_T)]+\lambda\,\mathbb{E}[\mathrm{Cost}(T)]$ with measurable surrogates, (ii) trace the

induced error–cost Pareto frontier for a family of admissible policies, and (iii) exhibit qualitative behaviors predicted by Thms. 1–5, including regimes in which ICL-only saturates under shift and regimes in which low-rank PEFT yields a predictable improvement as a function of r and s.

Benchmarks and task construction (cross-domain few-shot). We focus on cross-domain few-shot learning (CD-FSL) episodes in which q_{train} and q_{test} differ by domain, label space, or both. Concretely, we propose standard CD-FSL suites in vision (e.g., train on a source such as miniImageNet and test on targets such as CUB, Cars, Places, Plantae; or more severe shifts such as DomainNet), and, when applicable, analogous suites in language (domain-shifted classification and regression tasks with a shared backbone and episodic support/query splits). Each episode T provides S_T (size K) and Q_T, with N-way-K-shot sampling for classification and K-shot sampling for regression. We log both average risk and tail behavior across tasks (e.g., quantiles over $T \sim q_{\text{test}}$), since gating is intended to control per-task allocation rather than only mean performance.

Policies and ablations. We evaluate a family of policies aligned with the restricted class Π considered in Thm. 2. At minimum, we include:

1. **ICL-only:** choose $C_T \subseteq \mathcal{M}$ with $|C_T| \leq L$ (vary L), no parameter update ($\Delta_T = 0$).
2. **PEFT-only:** set $C_T = \emptyset$, fit a rank-r adapter for s steps on S_T (vary r, s).
3. **Hybrid (ours):** retrieve/select C_T and optionally fit Δ_T according to a gating statistic u_T and a threshold $\tau(\lambda)$.
4. **Full finetune (upper envelope):** update the full model (or full head) for a fixed compute budget, recorded as a high-cost reference point rather than a competitor under the same constraints.
5. **Oracle variants:** (a) oracle gating that chooses among ICL/PEFT/hybrid using query labels (infeasible, used only to upper-bound the achievable frontier), and (b) oracle retrieval that provides an idealized C_T (e.g., nearest neighbors in a privileged embedding) to isolate retrieval failures from adaptation failures.

Within each policy, we ablate (i) retrieval pool size m, (ii) demonstration selection (random, similarity-based, and greedy submodular surrogates), (iii) ordering of demonstrations in the context (random order, similarity order, diversity order), and (iv) adapter initialization (zero, pretrained default, meta-learned prior). These ablations separate the contributions of Retrieve($\cdot$), SelectDemos($\cdot$), and the adapter optimization routine.

Cost measurement and normalization. Our theoretical cost is $\mathrm{Cost}(T) = \alpha L + \beta r s + \gamma\, \mathrm{Retr}(L)$. Experimentally, we will report both (i) this proxy cost under user-chosen (α, β, γ) and (ii) direct system measurements: wall-clock latency, number of forward passes, number of backward passes, peak memory, and retrieval query time. The intent is not to enforce a single universal conversion between these units, but to verify that conclusions are stable across reasonable calibrations. For each λ, we compute the empirical objective

$$\widehat{J}_\lambda(\pi) \;:=\; \frac{1}{n}\sum_{j=1}^{n} \widehat{\mathcal{L}}_{T_j}(\hat{f}^{\pi}_{T_j}) \;+\; \lambda\, \frac{1}{n}\sum_{j=1}^{n} \widehat{\mathrm{Cost}}(T_j;\pi),$$

and we sweep λ to trace an empirical Pareto curve. We additionally report the achieved distribution of chosen actions (ICL-only vs PEFT-enabled) as a function of λ, which is a direct observable consequence of the gating mechanism.

Robustness under domain shift and label mapping. To connect to the shift parameter δ, we propose controlled stress tests in which shift magnitude is increased along interpretable axes. In vision, this includes corruptions and style shifts (blur, noise, color jitter, texture bias) and domain transfer (photographic to sketch/clipart). In language, we include temporal drift, topic drift, and instruction paraphrases that preserve semantics but alter surface form. For classification, we add a *label mapping shift*: retrieved demonstrations may come from a label set whose names or indices do not match the current task. We explicitly test policies under (i) correct mapping, (ii) permuted mapping, and (iii) ambiguous mapping where only textual class descriptions are provided. The hypothesis consistent with Thm. 3 is that, when the effective mapping or representation is shifted in a way not expressible as additional labeled examples in a fixed feature space, ICL-only performance saturates even as L grows, whereas PEFT can recover by altering the representation (subject to the rank constraint).

Failure cases aligned with lower bounds. We will include synthetic and semi-synthetic episodes designed to make the lower-bound mechanisms visible rather than merely implicit. For Thm. 3, we construct tasks where the discriminative direction lies largely in a shifted subspace (e.g., apply a hidden orthogonal transform A_T on embeddings or on a learned intermediate layer) such that the conditional distribution of $\phi(x)$ under the observed support is insufficient to identify the correct predictor without modifying the representation. We then evaluate ICL-only as L increases (including idealized C_T) to confirm the predicted plateau. For Thm. 4, we construct a family where the shift decomposes into $r^\star$ principal directions and measure excess risk as a function of adapter rank r, checking for an empirical $1/r$-type

decay once estimation error is controlled (e.g., by fixing K large enough or averaging across episodes).

Sensitivity to retrieval quality and ordering. Since the theory isolates retrieval as a resource with its own cost $\mathrm{Retr}(L)$ and does not assume it is perfect, we test the degradation of ICL utility under controlled retrieval noise: approximate nearest neighbors at varying recall, embedding quantization, and adversarially perturbed keys. We also examine ordering effects: even if C_T is fixed, the sequence presented to the model can change performance. We measure ordering sensitivity as a variance component of $\widehat{\mathcal{L}}_T$ over random permutations, and we test simple ordering rules (similarity-first, diversity-first) as low-cost interventions. A central diagnostic is whether the hybrid policy is *stable* to retrieval perturbations: when retrieval becomes unreliable, the gate should shift probability mass toward PEFT-enabled actions for the same λ, reflecting a rational substitution of resources.

What constitutes confirmation. We do not require exact constants from Thm. 1; rather, we seek the structural signatures: (i) ICL improves primarily through an estimation-like regime with diminishing returns in L, (ii) PEFT reduces a shift-like error component as rank/steps increase, and (iii) the hybrid gate yields a uniformly better empirical objective $\widehat{J}_\lambda$ across λ, approaching the oracle envelope while respecting measured costs. These experiments, taken together, are intended to show that the proposed competitive analysis is not merely a proof artifact, but a useful description of how adaptation resources should be allocated under realistic constraints.

11 Discussion and Future Directions

We view the preceding formulation as an attempt to make *adaptation* an object of algorithmic design rather than an implicit, model-specific convention. The central output of the theory is not a single best method, but a language for stating and testing claims of the form: under a task family q_{test}, an adaptation policy $\pi \in \Pi$ achieves a certain point on an error–cost Pareto frontier, where cost accounts for context length, PEFT computation, and retrieval overhead. This perspective suggests several implications for evaluation practice, several actionable heuristics for practitioners, and several open problems which are obscured when one reports only accuracy at a fixed prompt length or only accuracy at a fixed finetuning budget.

Toward unified evaluation protocols. The objective $\mathbb{E}[\mathcal{L}_T(\hat{f}_T)]+\lambda\,\mathbb{E}[\mathrm{Cost}(T)]$ compels us to report adaptation results as *curves* rather than single numbers. In particular, a policy that is superior at one λ may be dominated at another, and the hybrid gate is designed precisely to move along that curve

automatically as λ varies. We therefore recommend that benchmarks report: (i) an empirical Pareto frontier obtained by sweeping λ (or sweeping budgets B), (ii) the induced distribution over actions (ICL-only, PEFT-only, hybrid) as a function of λ, and (iii) a decomposition of cost into its constituents αL, βrs, and $\gamma \operatorname{Retr}(L)$, alongside wall-clock and memory measurements. These reports make it possible to compare methods even when system constants differ, because they permit re-weighting cost terms *post hoc*.

A second protocol implication is that we should evaluate robustness across *shift regimes*. Thm. 3 formalizes the phenomenon that, with a frozen ϕ, there exist shifts (captured abstractly by A_T) for which ICL-only saturates at an $\Omega(\delta^2)$ excess risk, regardless of L. Consequently, reporting performance only at mild shifts can systematically overstate the generality of ICL-based conclusions. It is more informative to report performance as a function of controlled shift severity (corruptions, domain transfer, label mapping ambiguity), and to include the failure modes where ICL is *provably* incapable of compensating for the shift. In such regimes, a successful method should not merely gain accuracy; it should reallocate resources toward representation change (via Δ_T) in a manner consistent with its cost model.

Finally, retrieval must be treated as a first-class experimental axis. Since $\operatorname{Retr}(L)$ enters the cost and since demo utility depends on retrieval quality, benchmark reports should include retrieval recall/latency tradeoffs, indexing/quantization settings, and the sensitivity of policies to retrieval perturbations. Without these measurements, one cannot distinguish a method that is intrinsically sample-efficient from a method that relies on an unusually favorable retrieval configuration.

Guidance for practitioners under budgets. The main prescriptive lesson is that one should not commit to either ICL or PEFT as a universal default; rather, one should implement a *resource allocation rule* that can substitute between them as conditions change. Concretely, given measured system coefficients (α, β, γ), one may treat λ as a dial reflecting latency or monetary constraints and deploy a gate based on an observable statistic u_T computed from S_T. When u_T indicates a small shift (e.g., support embeddings lie near known source prototypes and exhibit low dispersion), ICL-only with a modest L tends to be cost-effective because it reduces an estimation-like component without paying backprop cost. When u_T indicates substantial shift (or when retrieval is unreliable), it can be rational to spend βrs to reduce a shift-dominated component, consistent with the tradeoff in Thm. 1 and the tightness statement in Thm. 4.

In operational terms, we suggest the following workflow. First, calibrate $\operatorname{Cost}(T)$ by measuring the marginal latency and memory impact of: adding one more demonstration, increasing r by a small increment, and adding one more PEFT step s, as well as the retrieval overhead for the intended $\mathcal{M}$.

Second, choose a *target* λ (or B) corresponding to an application budget. Third, tune the gate (including u_T and $\tau(\lambda)$) on held-out tasks drawn from the anticipated deployment mixture. The quantity to tune is not accuracy alone but the empirical objective $\widehat{J}_\lambda$, since this is what will remain stable when deployment constraints change. Fourth, monitor the gate's action frequencies and failure cases: if the gate persistently chooses PEFT under apparently easy tasks, u_T is likely miscalibrated; if it persistently chooses ICL under hard shifts, then either u_T is underestimating shift or the chosen r, s are insufficient for the deployment δ.

Open questions beyond the linearized regime. Our analysis relies on a linear task model in an embedding space and a low-rank approximation of shift. While this abstraction yields crisp statements, several extensions are necessary for a complete account.

(i) *Beyond fixed ϕ and linearization:* one may seek analogues of Thm. 1 in regimes where adaptation changes ϕ in a nonlinear manner and where ICL cannot be faithfully represented as ridge regression in a fixed feature map. A plausible direction is to analyze ICL and PEFT through local linearization (e.g., tangent kernels) while explicitly tracking how A_T interacts with curvature and with the adapter parameterization.

(ii) *Nonparametric and structured shifts:* the operator A_T captures a broad class of domain shifts but does not encompass shifts that are inherently non-linear or label-conditional. It remains to understand when a low-rank correction is the right inductive bias and when one needs different structures (e.g., sparse, block-diagonal, or mixture-of-subspaces corrections). A related question is whether retrieval can be used not merely to reduce variance but to *identify* the relevant structure of the shift (thereby choosing r adaptively).

(iii) *Multimodal contexts and formatting constraints:* in multimodal systems, L is not simply a count of examples, and the relevant cost may be dominated by tokenization, image resolution, or cross-attention complexity. Moreover, the utility of demonstrations depends on formatting and interleaving across modalities. Developing a cost model that meaningfully predicts $\mathrm{Cost}(T)$ and a selection rule $\mathrm{SelectDemos}(\cdot)$ that remains approximately submodular under such constraints is open.

(iv) *Learning the policy class itself:* Thm. 2 is stated for a restricted Π. In practice, one may wish to enlarge Π to include policies that choose L, r, s continuously, policies that condition retrieval on queries, or policies that allocate budget across tasks in a streaming fashion. This points to online and bandit formulations in which λ and even (α, β, γ) are uncertain and must be estimated, while still respecting no-leakage constraints.

We expect that progress on these questions will require keeping the present discipline: separating statistical limitations (what can be learned from S_T and $\mathcal{M}$) from algorithmic limitations (NP-hard selection and con-

strained optimization), and reporting outcomes as error–cost tradeoffs rather than isolated accuracy points.