CARROT: A COST AWARE RATE OPTIMAL ROUTER

Seamus Somerstep[†□] **Felipe Maia Polo**[†□] **Allysson Flavio Melo de Oliveira** [‡∘]
Prattyush Mangal [‡] **Mírian Silva** [‡∘△] **Onkar Bhardwaj** [‡∘]
Mikhail Yurochkin[*‡∘] **Subha Maity**[*♠]

[□] Department of Statistics, University of Michigan
[‡] IBM Research [∘] MIT-IBM Watson AI Lab [△] Federal University of Minas Gerais
[♠] Department of Statistics and Actuarial Science, University of Waterloo

huggingface.co/CARROT-LLM-Routing

ABSTRACT

With the rapid growth in the number of Large Language Models (LLMs), there has been a recent interest in *LLM routing*, or directing queries to the cheapest LLM that can deliver a suitable response. Following this line of work, we introduce CARROT, a Cost AwaRe Rate Optimal rouTer that can select models based on any desired trade-off between performance and cost. Given a query, CARROT selects a model based on estimates of models' cost and performance. Its simplicity lends CARROT computational efficiency, while our theoretical analysis demonstrates minimax rate-optimality in its routing performance. Alongside CARROT, we also introduce the Smart Price-aware ROUTing (SPROUT) dataset to facilitate routing on a wide spectrum of queries with the latest state-of-the-art LLMs. Using SPROUT and prior benchmarks such as Routerbench and open-LLM-leaderboard-v2 we empirically validate CARROT's performance against several alternative routers.

1 INTRODUCTION

Large language models (LLMs) have demonstrated the capability to effectively address a diverse array of tasks across academic, industrial, and everyday settings (Minaee et al., 2024). This continued success has catalyzed the rapid development of new LLMs tailored for both general and specialized applications (Myrzakhan et al., 2024). While this offers practitioners increased flexibility, the vast number of available options may pose a daunting challenge in their real-world deployment. Particularly, determining the optimal LLM for a given query remains a significant challenge. In a perfect world, all queries can be routed to the most powerful model, but for many, this may quickly become prohibitively expensive.

A common approach to address this issue is *routing* (Shnitzer et al., 2023; Hu et al., 2024; Ong et al., 2024; Jain et al., 2023; Šakota et al., 2024). There are two paradigms of routing; *non-predictive* routers repeatedly call LLMs and evaluate the responses to select the best one for a given query. Examples include Fusion of Experts (FoE) (Wang et al., 2023), FrugalGPT (Chen et al., 2024), and techniques that cascade answers from weak to strong LLMs (Yue et al., 2024). The obvious disadvantage non-predictive routing is the required inference of many LLMS for all queries, even those that are not suitable for the task at hand. As a workaround, researchers have also considered *predictive routers*, which take LLM queries as inputs and output guesses at the most appropriate LLM. A key limitation of the prior literature on predictive routing is *the avoidance of the cost prediction problem*. In Shnitzer et al. (2023) only performance is considered. The methods RouteLLM (Ong et al., 2024) RoRF (Jain et al., 2023) incorporates model cost by creating binary routers that select between a large costly model and a cheap small model. While this does incorporate cost, we shall see that the reduced flexibility of binary routing leads to performance degradation in practice. Finally, in Šakota et al. (2024) cost prediction is considered, but the authors assume that cost may be inferred from available API pricing; this does not hold if model responses are open-ended, and thus may vary in length.

[†] primary contributors and corresponding authors (smrstep@umich.edu, maiapolo@umich.edu)
[*] lead investigators (mikhail.yurochkin@ibm.com, smaity@uwaterloo.ca)

To address these shortcomings, we develop CARROT: a Cost AwaRe Rate Optimal rouTer and collect the Smart Price-aware ROUTing (SPROUT) dataset. SPROUT is representative of both performance and cost of state-of-the-art LLMs across varying practical use-cases. Additionally, we use it to train the final version of CARROT. CARROT is designed to function as an estimate of *the oracle router*. As a concrete example of the oracle router, consider model performance and inference cost as the metrics of interest. First, based on their needs, a practitioner selects a convex combination of performance and cost as the model metric. Next, the oracle router takes in a query and produces the LLM that minimizes this specific model metric. To do this selection, the oracle router needs to perfectly predict the metrics of interest (cost, performance) for each model given any query. Inspired by this, CARROT utilizes a simple two-stage approach. We first attain an estimator for each of the metrics (*e.g.* cost and performance) for each model given a query, then we plug in these estimators to the formed risk function and select a model that minimizes the appropriate convex combination of the estimated metrics. Although CARROT utilizes a simple approach, it has three distinct advantages.

Computational efficiency: A practitioner may be interested in several possible convex combinations to understand the trade-offs across various metrics. In that case, learning a router per combination may not be practical. Our approach allows re-use of a plug-in estimate of the metrics for each combination, easing computational concerns. Furthermore, the metrics of interest may vary from practitioner to practitioner. CARROT allows the flexible addition of any metrics; one only needs to acquire estimators for them. As a demonstration of CARROT's efficiency, we utilize it to estimate the Pareto frontier of performance and cost trade-off on RouterBench[1](Hu et al., 2024), open-LLM-leaderboard-v2 [2] (Fourrier et al., 2024), and our new SPROUT dataset.

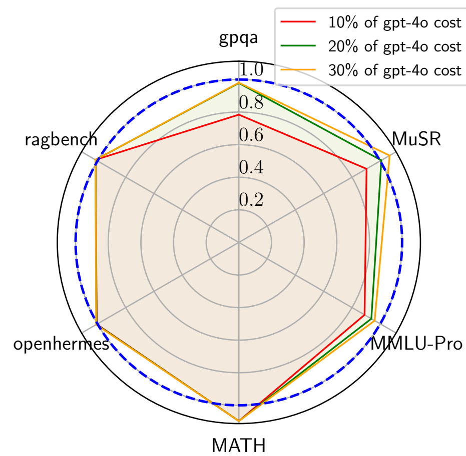

Figure 1: Percent of GPT-4o performance achieved by CARROT across datasets at various discounted costs, where the blue dotted line indicates similar (100%) performance to GPT-4o.

New dataset for routing: SPROUT covers 13 state-of-the-art language models (e.g. Llama-3-herd (Grattafiori et al., 2024), GPT-4o (Achiam et al., 2024), *etc.*) and approximately 45k prompts from 6 benchmarks covering RAG, science, reasoning, and GPT-4 generated user queries. For all models, we use zero-shot prompting and corresponding chat templates to represent practical use cases and collect input and output token counts to allow flexibility when studying cost-performance trade-offs. We defer further details about SPROUT and its advantages over other routing datasets to Section 4.1 and Appendix A.

As a sneak peek, in Figure 1, we present the ratio of CARROT's performance to GPT-4o's (Achiam et al., 2024) on several key benchmarks across diverse use cases represented in SPROUT. At 30% of the cost, CARROT matches or exceeds the performance of GPT-4o on each benchmark. Both SPROUT and the corresponding CARROT router will be publicly released.

Statistical efficiency Collecting adequate data to train a router is quite expensive as we need to generate and evaluate with every LLM for every query. Thus, statistical efficiency is an important consideration for any routing procedure. To investigate statistical efficiency, we connect the routing problem to multi-objective classification with a wide class of possible loss functions (this depends on the convex combination of model metrics on hand). Here, identities of the possible models act as the label space of the multi-class classification problem. Given this, one contribution of our work is to extend previous minimax studies in nonparametric classification (Audibert and Tsybakov, 2007) on two fronts: (1) with more than two classes, and (2) with general losses beyond $0/1$-loss. Both of these

[1]`https://huggingface.co/datasets/withmartian/routerbench`

[2]`https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`

extensions require us to introduce a generalized definition of margin (*cf.* eq. equation 3.3), which reduces to the usual margin definition as in Audibert and Tsybakov (2007) when the classification task is binary and the loss is $0/1$. Additionally, our work analyzes minimax optimality on the entire collection of oracle classifiers in classification problems with all possible trade-offs between multiple objectives, which, to the best of our knowledge, is a novel contribution.

For space purposes, we discuss how these contributions fit into the related literature in Appendix **??**.

1.1 Notation and preliminaries

To begin, let us introduce our notation. We have M pre-trained LLMs indexed as $m \in [M] = \{1, \dots, M\}$ and K metrics indexed as $k \in [K] = \{1, \dots, K\}$. We denote a generic input or query as $X \in \mathcal{X}$, where $\mathcal{X}$ is the space of inputs. Thus, for any input X, the metrics of interest are stored in a $M \times K$ matrix. We denote this matrix as $Y \in \mathbf{R}^{M \times K}$, whose (m, k)-th entry $[Y]_{m,k}$ is the metric value for obtaining a prediction from the m-th model evaluated with respect to k-th metric. For all metrics, we assume that a lower value is preferred. With this convention, we shall also refer to them as risks. For a probability distribution P in the sample space $\mathcal{X} \times \mathbf{R}^{M \times K}$ we assume that the training dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ is an iid sample from P.

For the probability distribution P defined on the space $\mathcal{X} \times \mathbf{R}^{M \times K}$, we denote the marginal distribution of X by P_X. Let us denote $\mathrm{supp}(\cdot)$ as the support of a probability distribution. Within the space $\mathbf{R}^d$, we denote Λ_d as the Lebesgue measure, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ as the ℓ_2 and ℓ_∞-norms, and $\mathcal{B}(x, r, \ell_2)$ and $\mathcal{B}(x, r, \ell_\infty)$ as closed balls of radius r and centered at x with respect to the ℓ_2 and ℓ_∞-norms.

2 Plug-in approach to routing

We will consider a convex combination of our K metrics with coefficients $\mu \in \Delta^{K-1} \triangleq \{(\mu_1, \dots, \mu_K) : \mu_k \geq 0, \sum_k \mu_k = 1\}$ and a generic point $(X, Y) \sim P$. The μ-th convex combination of the risks (or, μ-th risk) can be written as $Y\mu \in \mathbf{R}^M$, with the risk incurred for obtaining a prediction from the m-th model is

$$[Y\mu]_m = \textstyle\sum_{k=1}^K [Y]_{m,k} \mu_k \,.$$

We want to learn a predictive router $g : \mathcal{X} \to [M]$, that takes X as an input and predicts the index of the LLM to be used for inference. The average μ-th risk for using the router g is

$$\mathcal{R}_P(g, \mu) = \mathbf{E}\big[\textstyle\sum_{m=1}^M [Y\mu]_m \mathbb{I}\{g(X) = m\}\big] \,. \tag{2.1}$$

For a given μ let us refer to the minimizer $g_\mu^\star$ as an oracle router. Remember our objective: we would like to learn the oracle routers $g_\mu^\star$ at every value of μ. While one may minimize an empirical risk corresponding to $\mathcal{R}_P(g, \mu)$ to estimate the oracle router at a particular μ, this approach is not scalable, any small change in μ would require refitting a new router. Given this, we develop a plug-in approach which lets us estimate the oracle routers at every value of μ. The key intuition lies within an explicit form of the $g_\mu^\star$ that we provide in the next lemma.

Lemma 2.1. *Let us define $\Phi(x) = \mathbf{E}[Y \mid X = x]$ and $\eta_{\mu,m}(x) = \sum_{k=1}^K \mu_k [\Phi(x)]_{m,k}$. Then for any $\mu \in \Delta^{K-1}$ the oracle router that minimizes $\mathcal{R}_P(g, \mu)$ is*

$$\begin{aligned} g_\mu^\star(X) &= \arg\min\nolimits_m \, \eta_{\mu,m}(X) \\ &= \arg\min\nolimits_m \big\{ \textstyle\sum_{k=1}^K \mu_k [\Phi(X)]_{m,k} \big\} \,. \end{aligned} \tag{2.2}$$

The key conclusion of 2.1 is the expression $g_\mu^\star(X) = \arg\min_m \{\sum_{k=1}^K \mu_k [\Phi(X)]_{m,k}\}$. It suggests a straightforward approach to estimate $g_\mu^\star(X)$ at all values of μ. Namely, we only need to plug-in an estimate of $\Phi(X) = \mathbf{E}[Y \mid X]$ to the expression of $g_\mu^\star(X)$. Compared to minimizing empirical risk at different values of μ, this plug-in approach is more scalable if the practitioner plans on tuning μ. We summarize our approach in algorithm 1.

Algorithm 1 CARROT

Require: Dataset $\mathcal{D}_n$

1: Randomly split the dataset into training and test splits: $\mathcal{D}_n = \mathcal{D}_{\text{tr}} \cup \mathcal{D}_{\text{test}}$.
2: Learn an estimate $\hat{\Phi}(X)$ of $\Phi(X)$ using the $\mathcal{D}_{\text{tr}}$.
3: **for** $\mu \in \Delta^{K-1}$ **do**
4: Define $\widehat{g}_\mu(X) = \arg\min_m \widehat{\eta}_{\mu,m}(X)$ where $\widehat{\eta}_{\mu,m}(X) = \sum_{k=1}^K \mu_k [\widehat{\Phi}(X)]_{m,k}$. Break any tie within $\arg\min$ randomly.
5: Calculate $\widehat{\mathcal{R}}(\mu, k) = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(X,R)\in\mathcal{D}_{\text{test}}} [R]_{m,k}\mathbb{I}\{\widehat{g}_\mu(X) = m\}$
6: **end for**
7: **Return:** $\{\widehat{g}_\mu : \mu \in \Delta^{K-1}\}$ and $\{\widehat{\mathcal{R}}(\mu, k) : \mu \in \Delta^{K-1}, k \in [K]\}$.

3 Statistical efficiency of CARROT

In this section we establish that, under certain conditions, the plug-in approach to routing is minimax optimal. To show this, we follow two steps:

- First we establish an information theoretic lower bound on the sample complexity for learning the oracle routers (*cf.* Theorem 3.6).
- Next, establish an upper bound for the minimax risk of plug-in routers (*cf.* Theorem 3.9). We show that under sufficient conditions on the estimates of $\mathbf{E}[Y \mid X]$ the sample complexity in the upper bound matches the lower bound. Together, they imply the statistical efficiency of the plug-in approach.

We begin with a notational convention for $g_\mu^\star(X)$. If the minimum is attained at multiple m's, we consider $g_\mu^\star(X)$ as a subset of $[M]$. On the contrary, if the minimum is uniquely attained, then $g_\mu^\star(X)$ refers to both the index m_X where the minimum is attained and the singleton set $\{m_X\} \subset [M]$. The distinction should be clear from the context.

We also generalize slightly to the setting where the last K_2 metrics are known functions of X, *i.e.* for $m \in [M], k \in \{K - K_2 + 1, \dots K\}$ there exist known functions $f_{m,k} : \mathcal{X} \to \mathbf{R}$ such that $[Y]_{m,k} = f_{m,k}(X)$. Since $\mathbf{E}[[Y]_{m,k} \mid X] = f_{m,k}(X)$ are known for $k \geq K - K_2 + 1$ they don't need to be estimated.

3.1 Technical Assumptions

The technical assumptions of our minimax study are closely related to those in investigations of non-parametric binary classification problems with $0/1$ loss functions, *e.g.* Cai and Wei (2019); Kpotufe and Martinet (2018); Maity et al. (2022); Audibert and Tsybakov (2007). In fact, our setting generalizes the classification settings considered in these papers on multiple fronts: (i) we allow for general loss functions, (ii) we allow for more than two classes, and (iii) we allow for multiple objectives.

To clarify this, we discuss how binary classification is a special case of our routing problem.

Example 3.1 (Binary classification with $0/1$-loss)**.** *Consider a binary classification setting with $0/1$-loss: we have the pairs $(X, Z) \in \mathcal{X} \times \{0, 1\}$ and we want to learn a classifier $h : \mathcal{X} \to \{0, 1\}$ to predict Z using X. This is a special case of our setting with $M = 2$ and $K = 1$, where for $m \in \{0, 1\}$ the $[Y]_{m,1} = \mathbb{I}\{Z \neq m\}$. Then the risk for the classifier h, which can also be thought of as a router, is*

$$\begin{aligned}\mathcal{R}_P(h) &= \mathbf{E}\big[\textstyle\sum_{m\in\{0,1\}}[Y]_{m,1}\mathbb{I}\{h(X) = m\}\big] \\ &= \mathbf{E}\big[\mathbb{I}\{h(X) \neq Z\}\big]\,,\end{aligned}$$

the standard misclassification risk for binary classification.

We assume that $\text{supp}(P_X)$ is a compact set in $\mathbf{R}^d$. This is a standard assumption in minimax investigations for non-parametric classification problems (Audibert and Tsybakov, 2007; Cai and Wei, 2019; Kpotufe and Martinet, 2018; Maity et al., 2022). Next, we place Hölder smoothness conditions on the functions $\Phi^\star_m$. This controls the difficulty of their estimation. For a tuple $s = (s_1, \ldots, s_d) \in (\mathbf{N} \cup \{0\})^d$ of d non-negative integers define $|s| = \sum_{j=1}^d s_j$ and for a function $\phi : \mathbf{R}^d \to \mathbf{R}$ and $x = (x_1, \ldots, x_d) \in \mathbf{R}^d$ define the differential operator:

$$D_s(\phi, x) = \tfrac{\partial^{|s|}\phi(x)}{\partial x_1^{s_1} \ldots \partial x_d^{s_d}}\,, \tag{3.1}$$

assuming that such a derivative exists. Using this differential operator we now define the Hölder smoothness condition:

Definition 3.2 (Hölder smoothness)**.** *For $\beta, K_\beta > 0$ we say that $\phi : \mathbf{R}^d \to \mathbf{R}$ is (β, K_β)-Hölder smooth on a set $A \subset \mathbf{R}^d$ if it is $\lfloor \beta \rfloor$-times continuously differentiable on A and for any $x, y \in A$*

$$|\phi(y) - \phi_x^{(\lfloor \beta \rfloor)}(y)| \le K_\beta \|x - y\|_2^\beta\,, \tag{3.2}$$

where $\phi_x^{(\lfloor \beta \rfloor)}(y) = \sum_{|s| \le \lfloor \beta \rfloor} D_s(\phi, x)\{\prod_{j=1}^d (y_j - x_j)^{s_j}\}$ is the $\lfloor \beta \rfloor$-order Taylor polynomial approximation of $\phi(y)$ around x.

With this definition, we assume the following:

Assumption 3.3. *For $m \in [M]$ and $k \in [K_1]$ the $[\Phi(X)]_{m,k}$ is $(\gamma_k, K_{\gamma,k})$-Hölder smooth.*

This smoothness parameter will appear in the sample complexity of our plug-in router. Since the $[\Phi(X)]_{m,k}$ are known for $k \ge K_1 + 1$ we do not require any smoothness assumptions on them.

Next, we introduce *margin condition*, which quantifies the difficulty in learning the oracle router. For a given μ define the margin as the difference between the minimum and second minimum of the risk values:

$$\Delta_\mu(x) = \begin{cases} \min\limits_{m \notin g_\mu(x)} \eta_{\mu,m}(x) - \min\limits_m \eta_{\mu,m}(x) & \text{if } g^\star_\mu(x) \ne [M] \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

Our definition of a margin generalizes the usual definition of the margin considered for binary classification with $0/1$ loss in Audibert and Tsybakov (2007). Recall the binary classification example in 3.1, in which case, $[\Phi(X)]_{m,1} = P(Z \ne m \mid X)$. Since $K = 1$ we have $\eta_{\mu,m}(X) = P(Z \ne m \mid X)$, which further implies $\eta_{\mu,0}(X) + \eta_{\mu,1}(X) = 1$. Thus for binary classification with $0/1$ loss, our definition of margin simplifies to

$$\min_{m \notin g^\star_\mu(x)} \eta_{\mu,m}(x) - \min_m \eta_{\mu,m}(x) = |\eta_{\mu,1}(X) - \eta_{\mu,0}(X)| = 2|\eta_{\mu,0}(X) - 1/2|\,,$$

which is a constant times the margin $|P(Y = 1 \mid X) - 1/2| = |\eta_{\mu,0}(X) - 1/2|$ in Audibert and Tsybakov (2007).

Clearly, the margin determines the difficulty in learning the oracle router. A query X with a small margin gap is difficult to route, because to have the same prediction as the oracle, *i.e.* $\arg\min_m \hat\eta_{\mu,m}(X) = \arg\min_m \eta^\star_{\mu,m}(X)$ we need to estimate $\eta^\star_{\mu,m}(X)$ with high precision. In the following assumption, we control the probability of drawing these "difficult to route" queries.

Assumption 3.4 (Margin condition)**.** *For $\alpha, K_\alpha > 0$ and any $t > 0$ the margin Δ_μ equation 3.3 satisfies:*

$$P_X\big\{0 < \Delta_\mu(X) \le t\big\} \le K_\alpha t^\alpha\,. \tag{3.4}$$

Following Audibert and Tsybakov (2007), we focus on the cases where $\alpha < d$ and for every k the $\alpha\gamma_k < d$. This helps to avoid trivial cases where routing decisions are constant over P_X for some μ. Next, we assume that P_X has a density p_X that satisfies a strong density condition described below.

Assumption 3.5 (Strong density condition)**.** *Fix constants $c_0, r_0 > 0$ and $0 \le \mu_{\min} \le \mu_{\max} < \infty$. We say P_X satisfies the strong density condition if its support is a compact (c_0, r_0)-regular set and it has density p_X which is bounded: $\mu_{\min} \le p_X(x) \le \mu_{\max}$ for all x within $\operatorname{supp}(P_X)$. A set $A \subset \mathbf{R}^d$ is (c_0, r_0)-regular if it is Lebesgue measurable and for any $0 < r \le r_0,\ x \in A$ it satisfies*

$$\Lambda_d(A \cap \mathcal{B}(x, r, \ell_2)) \ge c_0 \Lambda_d(\mathcal{B}(x, r, \ell_2)). \tag{3.5}$$

This is another standard assumption required for minimax rate studies in nonparametric classification problems (Audibert and Tsybakov, 2007; Cai and Wei, 2019). All together, we define $\mathcal{P}(c_0, r_0, \mu_{\min}, \mu_{\max}, \beta_{m,k}, K_{\beta,m,k}, \alpha, K_\alpha)$, or simply $\mathcal{P}$, as the class of probabilities P defined on the space $\mathcal{X} \times \mathcal{Y}$ for which P_X is compactly supported and satisfies the strong density assumption 3.5 with parameters $(c_0, r_0, \mu_{\min}, \mu_{\max})$, and the Hölder smoothness assumption 3.3 and the (α, K_α)-margin condition in Assumption 3.4 hold. We shall establish our minimax rate of convergence within this probability class.

3.2 The lower bound

Rather than the actual risk $\mathcal{R}_P(\mu, g)$, we establish a lower bound on the excess risk:

$$\mathcal{E}_P(\mu, g) = \mathcal{R}_P(\mu, g) - \mathcal{R}_P(\mu, g^\star_\mu)\,, \tag{3.6}$$

that compares the risk of a proposed router to the oracle one. We denote $\Gamma = \{g : \mathcal{X} \to [M]\}$ as the class of all routers. For an $n \in \mathbf{N}$ we refer to the map $A_n : \mathcal{Z}^n \to \Gamma$, which takes the dataset $\mathcal{D}_n$ as an input and produces a router $A_n(\mathcal{D}_n) : \mathcal{X} \to [M]$, as an algorithm. Finally, call the class of all algorithms that operate on $\mathcal{D}_n$ as $\mathcal{A}_n$. The following Theorem describes a lower bound on the minimax risk for any such algorithm A_n.

Theorem 3.6. *For an $n \geq 1$ and $A_n \in \mathcal{A}_n$ define $\mathcal{E}_P(\mu, A_n) = \mathbf{E}_{\mathcal{D}_n}\big[\mathcal{E}_P\big(\mu, A_n(\mathcal{D}_n)\big)\big]$ as the excess risk of an algorithm A_n. There exists a constant $c > 0$ that is independent of both n and μ such that for any $n \geq 1$ and $\mu \in \Delta^{K-1}$ we have the lower bound*

$$\min_{A_n \in \mathcal{A}_n} \max_{P \in \mathcal{P}} \; \mathcal{E}_P(\mu, A_n) \geq c\big\{ \textstyle\sum_{k=1}^{K_1} \mu_k n^{-\frac{\gamma_k}{2\gamma_k + d}} \big\}^{1+\alpha}\,. \tag{3.7}$$

This result is a generalization of that in Audibert and Tsybakov (2007), which considers binary classification.

Remark 3.7. *Consider the binary classification in Example 3.1. Since $K = 1$, the lower bound simplifies to $\mathcal{O}(n^{-\gamma_1(1+\alpha)/2\gamma_1+d})$, which matches with the rate in Audibert and Tsybakov (2007, Theorem 3.5). Beyond $0/1$ loss, our lower bound also establishes that the rate remains identical for other classification loss functions as well.*

3.3 The upper bound

Next, we show that if algorithm the A_n corresponds to CARROT, the performance of $\hat{g}_\mu$ matches the lower bound in Theorem 3.6 (*cf.* equation 3.7). En-route to attaining $\hat{g}_\mu$, we need an estimate $\widehat{\Phi}(X)$ of $\Phi(X) = \mathbf{E}_P[Y \mid X]$. Our strategy will consist of two steps:

- First, we establish an upper bound on the rate of convergence for excess risk equation 3.6 for the plug-in router in terms of the rate of convergence for $\widehat{\Phi}(X)$.
- Then we discuss the desired rate of convergence in $\widehat{\Phi}(X)$ so that the upper bound has the identical rate of convergence to the lower bound equation 3.7. Later in Appendix C.1 we provide an estimate $\widehat{\Phi}(X)$ that has the required convergence rate.

These two steps, together with the lower bound in equation 3.7 establish that our plug-in router achieves the best possible rate of convergence in excess risk.

We begin with an assumption that specifies a rate of convergence for $[\widehat{\Phi}(X)]_{m,k}$.

Assumption 3.8. *For some constants $\rho_1, \rho_2 > 0$ and any $n \geq 1$ and $t > 0$ and almost all X with respect to the distribution P_X we have the following concentration bound:*

$$\max_{P \in \mathcal{P}} P\big\{ \max_{m,k} a_{k,n}^{-1} \big|[\widehat{\Phi}(X)]_{m,k} - [\Phi(X)]_{m,k}\big| \geq t \big\} \leq \rho_1 \exp\big(-\rho_2 t^2\big)\,, \tag{3.8}$$

where for each k the $\{a_{k,n}; n \geq 1\} \subset (0, \infty)$ is a sequence that decreases to zero.

Using this high-level assumption, in the next theorem, we establish an upper bound on the minimax excess risk for CARROT that depends on both $a_{k,n}$ and μ.

Theorem 3.9 (Upper bound). *Assume 3.8. If all the $P \in \mathcal{P}$ satisfy the margin condition 3.4 with the parameters (α, K_α) then there exists a $K > 0$ such that for any $n \geq 1$ and $\mu \in \Delta^{K-1}$ the excess risk for the router $\widehat{g}_\mu$ in Algorithm 1 is upper bounded as*

$$\max_{P \in \mathcal{P}} \mathbf{E}_{\mathcal{D}_n}\big[\mathcal{E}_P(\widehat{g}_\lambda, \lambda)\big] \leq K\big\{ \textstyle\sum_{k=1}^{K_1} \mu_k a_{k,n} \big\}^{1+\alpha} . \tag{3.9}$$

Remark 3.10 (Rate efficient routers). *When $a_{k,n} = n^{-\gamma_k/(2\gamma_k+d)}$ the upper bound in Theorem 3.9 has the $\mathcal{O}(\{\sum_{k=1}^{K_1} \mu_k n^{-\gamma_k/(2\gamma_k+d)}\}^{1+\alpha})$-rate, which is identical to the rate in the lower bound (cf. Theorem 3.6), suggesting that the minimax optimal rate of convergence for the routing problem is*

$$\min_{A_n \in \mathcal{A}_n} \max_{P \in \mathcal{P}} \ \mathcal{E}_P(A_n, \lambda) \asymp \mathcal{O}\big(\big\{ \textstyle\sum_{k=1}^{K_1} \mu_k n^{-\frac{\gamma_k}{2\gamma_k+d}} \big\}^{1+\alpha}\big) . \tag{3.10}$$

Following this, we conclude: When $a_{k,n} = n^{-\gamma_k/(2\gamma_k+d)}$ the plug-in approach in Algorithm 1, in addition to being computationally efficient, is also minimax rate optimal.

An example of an estimator $\widehat{\Phi}$ that meets the needed conditions for $a_{k,n} = n^{-\gamma_k/(2\gamma_k+d)}$ to hold is described in Appendix C.1.

4 Routing in benchmark case-studies

We use CARROT (Algorithm 1) to perform routing on several benchmark datasets.

4.1 Datasets

RouterBench: RouterBench (Hu et al., 2024) is a benchmark dataset for routing tasks consisting of approximately 30k prompts and responses from eleven ($M = 11$) different LLMs. The data includes prompts from 8 benchmarks covering commonsense reasoning, knowledge-based understanding, conversation, math, and coding.

Open LLM leaderboard: The Open LLM leaderboard v2[3] (Fourrier et al., 2024) is an open-source benchmarking platform that comprises responses and evaluations of a collection of LLMs on six benchmarks comprising a diverse collection of tasks.

SPROUT: We introduce (and evaluate CARROT on) SPROUT, a large and diverse dataset designed for training and evaluating routers. SPROUT integrates 13 state-of-the-art language models and prompts from 6 benchmarks, including GPQA (Rein et al., 2023), MuSR (Sprague et al., 2024), MMLU-Pro (Wang et al., 2024), MATH (Hendrycks et al., 2021b), OpenHermes (Teknium, 2023), and RAGBench (Friel et al., 2025). Compared to existing routing benchmarks such as RouterBench, SPROUT offers several key advantages:

1. SPROUT encompasses a highly diverse set of questions, including instruction queries.
2. Unlike previous benchmarks, it does not rely on few-shot prompting and utilizes chat templates appropriate for each model, making it more representative of real-world use cases.
3. It leverages LLaMa-3.1-70b-Instruct (Grattafiori et al., 2024) to evaluate LLM responses against the ground truth, similarly to Ni et al. (2024). This is crucial for evaluating on open-ended instruction queries as well as mitigating errors associated with traditional automatic evaluation methods like exact match.
4. We provide input and output token counts for each LLM-prompt pair, enabling flexibility when conducting cost-aware analysis.

We will open-source SPROUT and a platform that allows practitioners to extend SPROUT by adding new queries and seamlessly evaluating state-of-the-art models on them. For further details, please refer to Appendix A.

[3] `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`

4.2 ESTIMATING THE ORACLE ROUTER

CARROT requires an estimate for the function $\Phi^\star_m(X) = \mathbf{E}_P[Y_m \mid X]$. In our benchmark tasks, Y_m is 2-dimensional, consisting of model performance measured as accuracy and model cost measured in dollars. In the RouterBench and SPROUT datasets the cost is treated as unknown, while in the Open LLM leaderboard v2, most evaluations are likelihood-based, thus cost is essentially the length of the input, i.e., a known function. In all cases, the accuracy $Y_{\text{acc},m}$ is binary, and thus we can view its estimation as a binary classification problem, where our objective is to predict the probability that m-th model will answer the question X correctly, *i.e.* $P_m(X) = P(Y_{\text{acc},m} = 1 \mid X))\, Y_{\text{acc},m}$. Using this intuition, we train several multi-label classification models $\widehat{P} : \mathcal{X} \to [0,1]^M$ on a training data split consisting of 80% of the full dataset, where the m-th coordinate of $\widehat{P}(X)$ is the predicted probability that m-th model accurately answers the question X. To train $\widehat{P}$ we consider two procedures:

1. **CARROT (KNN):** We embed the model inputs using the text-embedding-3-small model from OpenAI (OpenAI, 2023). On these text embeddings, we train a multi-label K-nearest-neighbors (KNN) classifier.
2. **CARROT (Roberta):** We fine-tune the pre-trained weights of the roberta-base[4] architecture. In order to enhance efficiency, across m we allow $\hat{P}$ to share the same network parameters, except for the final classification layer.

As mentioned above, in the RouterBench and SPROUT task the cost remains to be estimated. We train multi-label regression models $\widehat{C} : \mathcal{X} \to \mathbf{R}^M$, where $\hat{C}_m(X) = \mathbb{E}[Y_{\text{cost},m}|X]$ is the estimated cost of calling model m for query X. In the case of SPROUT we actually estimate the input and output token count, and convert this to a cost using collected pricing numbers (see Table 2). Depending on the technique used for performance estimation, we use either the roberta-base or text-embedder plus KNN strategy outlined above for fitting $\hat{C}$ (note that the models may be altered for a multi-output regression task, but otherwise remain identical). In the case of Open LLM leaderboard v2, we compute the cost of input X by calculating its length measured as the number of input tokens and retrieving the price per token from TogetherAI (see Table 3 for prices per 1M of input tokens).

Having estimated the $\widehat{\Phi}_m$'s, we use them in Algorithm 1 to estimate the oracle routers and predict their performances vs costs on the remaining 20% test split of the dataset. These cost-accuracy tradeoff curves are provided in Figure 2 (left for RouterBench), Figure 2 (right Open LLM leaderboard v2), and Figure 3 (left for SPROUT) along with the accuracies and costs of the individual LLMs within the dataset.

4.3 BASELINE METHODS

Ong et al. (2024) (RouteLLM) proposes a collection of methods for learning binary routers from preference data (data consisting of queries q and labels $l_{i,j}$ indicating a winner between model i and j). While the usage of preference data is slightly different from ours, we implement their methods on our data by creating pseudo-preference data between two models. In particular, we select a costly and high-performing model and a cheaper model and say the costly model wins if and only if it is correct while the cheaper model is incorrect. On this pseudo preference data, we fit two methods from Ong et al. (2024) for learning win probabilities between expensive and cheap models: the first is a matrix factorization method, called **RouteLLM (MF)**, while the second uses fine-tuned roberta-base, called **RouteLLM (Roberta)**. A follow-up method to these is Routing on Random Forests (RoRF) from Not-Diamond (Jain et al., 2023), referred to as **Not-Diamond RoRF**. This method uses a text-embedder and random forest model to predict the win probability; we provide a comparison to this method with the text-embedding-3-small embedder from OpenAI. As in Jain et al. (2023), we use a slightly different procedure to construct the preference data; a label can take one of four possible values (one for each combination of correct/incorrect from each model), and the costly model is favored if either both models are wrong or the cheaper model is incorrect while the expensive model is correct. For RouterBench, we consider GPT-4 and mixtral 8x7b to be the costly and cheaper models while for the Open LLM Leaderboard, we use Qwen2.5 72b and Qwen2.5 7b.

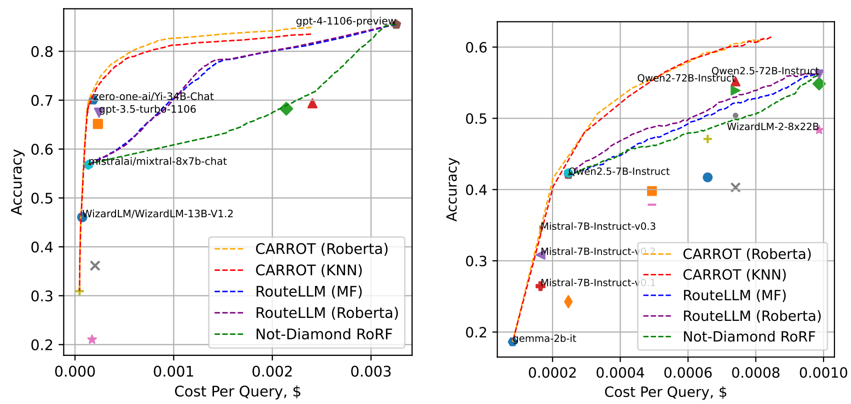

Figure 2: Left: Routerbench Right: Open LLM leaderboard v2

4.4 RESULTS

Performance against baselines: In RouterBench we were unable to achieve significantly better accuracy than GPT-4; however, we were able to greatly reduce the prediction cost. A direct comparison between routers and GPT-4 with respect to average accuracy versus cost is provided in Figure 2 (left), where we see that routers can achieve an accuracy similar to GPT-4 at half the cost, while achieving 95% of the accuracy at only 20% of the cost. On the other hand, we showed that CARROT can outperform the best model (Qwen2-72B) by a large margin in Open LLM leaderboard v2 (see the right side of Figure 2). In both datasets, we see that our routers significantly outperform the binary routers of Ong et al. (2024) and Jain et al. (2023). This is due to the fact that we route to *all possible models*, which increases the accuracy coverage and decreases the cost of the cheapest accurate model for a given query.

To verify our theory, we also compare to the ERM router, which directly minimizes ERM risk of equation 2.1 for a particular combination of accuracy and cost metrics. CARROT matches the ERM router performance, as demonstrated in Appendix Figures 5 and 6, verifying Lemma 2.1.

Performance on SPROUT: The performance of CARROT on SPROUT is illustrated in Figure 3 (left); a trend similar to our findings for RouterBench and the Open-LLM Leaderboard-v2 emerges. Moreover, Figure 1 shows that CARROT can achieve similar or better performance of GPT-4o with a fraction of the cost.

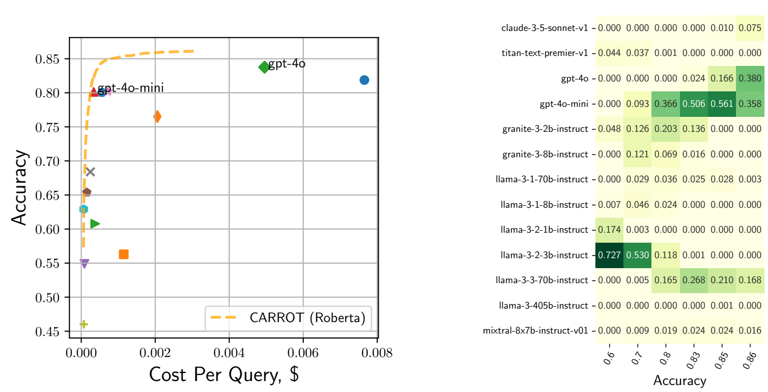

Figure 3: Left: Test Performance on SPROUT Right: Model selection proportions on SPROUT

[4] `https://huggingface.co/FacebookAI/roberta-base`

How CARROT chooses models: Figure 3 (right) presents the distribution of selected models in SPROUT across some (average cost, average accuracy) pairs in the test split. As we move from left to right in Figure 3 (right), the selection strategy gradually shifts from prioritizing cost efficiency with smaller models such as llama-3-2-1b-instruct and llama-3-2-3b-instruct to favoring more capable and expensive models like llama-3-3-70b-instruct, gpt-4o, and gpt-4o-mini.

5 Discussion

We introduced CARROT, a plug-in based router that is both computationally and statistically efficient. The computational efficiency stems from the requirement of merely calculating the plug-in estimators (see Algorithm 1) to perform routing Since collecting adequate data for router training might be challenging, we investigate CARROT's statistical efficiency in routing through a minimax rate study. To establish the statistical efficiency of CARROT, we have provided an information-theoretic lower bound on the excess risk of any router in Theorem 3.6 and corresponding upper bound for CARROT in Theorem 3.9. To ensure a broad scope for CARROT to a diverse set of queries *and* the latest state-of-the-art LLMs, we also introduced the SPROUT dataset.

Our routing and data approach is designed to be forward-looking. CARROT can incorporate many metrics besides performance and cost; an important next step is to explore which other metrics can improve LLM-based decision-making in practice. A related future goal is to benchmark our SPROUT-trained router on enterprise use cases like the Domain Intelligence Benchmark Suite (DIBS)[5] to locate areas of improvement needed in our data.

References

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report, 2024.

J.-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *Annals of Statistics*, 35 (2):608–633, Apr. 2007. ISSN 0090-5364, 2168-8966. doi: 10.1214/009053606000001217.

T. T. Cai and H. Wei. Transfer Learning for Nonparametric Classification: Minimax Rate and Adaptive Classifier. *arXiv:1906.02903 [cs, math, stat]*, June 2019.

V. Castelli, R. Chakravarti, S. Dana, A. Ferritto, R. Florian, M. Franz, D. Garg, D. Khandelwal, S. McCarley, M. McCawley, M. Nasr, L. Pan, C. Pendus, J. Pitrelli, S. Pujar, S. Roukos, A. Sakrajda, A. Sil, R. Uceda-Sosa, T. Ward, and R. Zhang. The TechQA dataset. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1269–1278, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.117. URL https://aclanthology.org/2020.acl-main.117.

L. Chen, M. Zaharia, and J. Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=cSimKw5p6R.

Z. Chen, W. Chen, C. Smiley, S. Shah, I. Borova, D. Langdon, R. Moussa, M. Beane, T.-H. Huang, B. Routledge, and W. Y. Wang. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711. Association for Computational Linguistics, Nov. 2021. doi: 10.18653/v1/2021.emnlp-main.300. URL https://aclanthology.org/2021.emnlp-main.300.

C. Fourrier, N. Habib, A. Lozovskaya, K. Szafer, and T. Wolf. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2024.

R. Friel, M. Belyi, and A. Sanyal. Ragbench: Explainable benchmark for retrieval-augmented generation systems, 2025. URL https://arxiv.org/abs/2407.11005.

[5]https://www.databricks.com/blog/benchmarking-domain-intelligence

A. Grattafiori, A. Dubey, and Others. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

D. Hendrycks, C. Burns, A. Chen, and S. Ball. Cuad: An expert-annotated nlp dataset for legal contract review. *NeurIPS*, 2021a.

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset, 2021b. URL https://arxiv.org/abs/2103.03874.

Q. J. Hu, J. Bieker, X. Li, N. Jiang, B. Keigwin, G. Ranganath, K. Keutzer, and S. K. Upadhyay. Routerbench: A benchmark for multi-LLM routing system. In *Agentic Markets Workshop at ICML 2024*, 2024. URL https://openreview.net/forum?id=IVXmV8Uxwh.

D. Jain, T.-Y. Tung, and T. H. Kofman. Rorf: Routing on random forests, 2023. URL https://www.notdiamond.ai/blog/rorf. Accessed: 2025-01-02.

Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577. Association for Computational Linguistics, Nov. 2019. doi: 10.18653/v1/D19-1259. URL https://aclanthology.org/D19-1259.

E. Kamalloo, A. Jafari, X. Zhang, N. Thakur, and J. Lin. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution, 2023.

S. Kpotufe and G. Martinet. Marginal Singularity, and the Benefits of Labels in Covariate-Shift. *arXiv:1803.01833 [cs, stat]*, Mar. 2018.

S. Maity, Y. Sun, and M. Banerjee. Minimax optimal approaches to the label shift problem in non-parametric settings. *Journal of Machine Learning Research*, 23(346):1–45, 2022.

C. Malaviya, S. Lee, S. Chen, E. Sieber, M. Yatskar, and D. Roth. Expertqa: Expert-curated questions and attributed answers, 2024.

S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

T. Möller, A. Reina, R. Jayakumar, and M. Pietsch. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.nlpcovid19-acl.18.

A. Myrzakhan, S. M. Bsharat, and Z. Shen. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena. *arXiv preprint arXiv:2406.07545*, 2024.

A. Nandy, S. Sharma, S. Maddhashiya, K. Sachdeva, P. Goyal, and N. Ganguly. Question answering over electronic devices: A new benchmark dataset and a multi-task learning based QA framework. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4600–4609, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.392. URL https://aclanthology.org/2021.findings-emnlp.392.

T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng. MS MARCO: A human-generated MAchine reading COmprehension dataset, 2017. URL https://openreview.net/forum?id=Hk1iOLcle.

J. Ni, F. Xue, X. Yue, Y. Deng, M. Shah, K. Jain, G. Neubig, and Y. You. Mixeval: Deriving wisdom of the crowd from llm benchmark mixtures. *arXiv preprint arXiv:2406.06565*, 2024.

I. Ong, A. Almahairi, V. Wu, W.-L. Chiang, T. Wu, J. E. Gonzalez, M. W. Kadous, and I. Stoica. Routellm: Learning to route llms with preference data, 2024. URL https://arxiv.org/abs/2406.18665.

OpenAI. Openai text-embedding-3-small model, 2023. URL https://platform.openai.com/docs/guides/embeddings/. Accessed: 2025-01-02.

D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark, 2023. URL https://arxiv.org/abs/2311.12022.

M. Sadat, Z. Zhou, L. Lange, J. Araki, A. Gundroo, B. Wang, R. Menon, M. Parvez, and Z. Feng. Delucionqa: Detecting hallucinations in domain-specific question answering. pages 822–835, 01 2023. doi: 10.18653/v1/2023.findings-emnlp.59.

T. Shnitzer, A. Ou, M. Silva, K. Soule, Y. Sun, J. Solomon, N. Thompson, and M. Yurochkin. Large language model routing with benchmark datasets. *arXiv preprint arXiv:2309.15789*, 2023.

Z. Sprague, X. Ye, K. Bostrom, S. Chaudhuri, and G. Durrett. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. *arXiv preprint arXiv:2310.16049*, 2023.

Z. Sprague, X. Ye, K. Bostrom, S. Chaudhuri, and G. Durrett. Musr: Testing the limits of chain-of-thought with multistep soft reasoning, 2024. URL https://arxiv.org/abs/2310.16049.

Teknium. Openhermes 2.5, 2023. URL https://huggingface.co/datasets/teknium/OpenHermes-2.5. Accessed: 2025-01-30.

H. Wang, F. M. Polo, Y. Sun, S. Kundu, E. Xing, and M. Yurochkin. Fusing models with complementary expertise. *arXiv preprint arXiv:2310.01542*, 2023.

Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, and W. Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, 2024. URL https://arxiv.org/abs/2406.01574.

Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

M. Yue, J. Zhao, M. Zhang, L. Du, and Z. Yao. Large language model cascades with mixture of thoughts representations for cost-efficient reasoning, 2024. URL https://arxiv.org/abs/2310.03094.

F. Zhu, W. Lei, Y. Huang, C. Wang, S. Zhang, J. Lv, F. Feng, and T.-S. Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.254. URL https://aclanthology.org/2021.acl-long.254.

M. Šakota, M. Peyrard, and R. West. Fly-swat or cannon? cost-effective language model choice via meta-modeling. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM ’24, page 606–615. ACM, Mar. 2024. doi: 10.1145/3616855.3635825. URL http://dx.doi.org/10.1145/3616855.3635825.

A SPROUT CONSTRUCTION DETAILS AND PLOTS

In this section, we discuss data details for SPROUT. SPROUT will be released on HuggingFace hub as a HuggingFace datasets object. For convenience, the data is pre-divided into train, validation, and test splits. Consider the training set as an example; the features of this split are

```

features = ['key', 'dataset', 'dataset level', 'dataset idx', 'prompt',
            'golden answer', 'aws-claude-3-5-sonnet-v1',
            'titan-text-premier-v1', 'openai-gpt-4o',
            'openai-gpt-4o-mini', 'granite-3-2b-instruct',
            'granite-3-8b', 'llama-3-1-70b-instruct',
            'llama-3-1-8b-instruct', 'llama-3-2-1b-instruct',
            'llama-3-2-3b-instruct', 'llama-3-3-70b-instruct',
            'llama-3-405b-instruct', 'mixtral-8x7b-instruct-v01']
```

Each key corresponds to another list. `"prompt"` contains the model queries, the `"dataset"` list indicates which sub-task a given query falls in (*cf.* Table 1 for info), and `golden answer` contains a desirable response for each query. Finally, the model keys each correspond to a list of dictionaries that contains further information on the responses of that model. The important keys in each dictionary of the list are `["num input tokens", "num output tokens", "response", "score"]`. They contain the number of input tokens for a query, the number of output tokens a model gives in response to a query, the actual response of the model, and finally the score that the judge provides for the response (using the corresponding `golden answer` entry). The conversion of token count to cost is given in Table 2 and additional details on the judging process are described in Section A.2.

A.1 SPROUT INGREDIENTS

Table 1 gives the benchmark ingredients for SPROUT. Namely, we use the MATH Lvl 1-5 (Hendrycks et al., 2021b), MMLU-PRO (Wang et al., 2024), GPQA (Rein et al., 2023), MUSR (Sprague et al., 2023), RAGBench (Friel et al., 2025), and openhermes (Teknium, 2023) datasets. These six benchmarks are varied and designed to simulate real-world scenarios where LLMs encounter a wide range of prompts. MATH focuses solely on mathematical word problems, whereas MMLU-PRO and GPQA include both mathematical and advanced science questions. MuSR serves as a benchmark for assessing multistep soft reasoning tasks framed within natural language narratives. RAGBench is a retrieval augmented generation (RAG) benchmark dataset collected from Question-Answer (QA) datasets (CovidQA (Möller et al., 2020), PubmedQA (Jin et al., 2019), HotpotQA (Yang et al., 2018), MS Marco (Nguyen et al., 2017), CUAD (Hendrycks et al., 2021a), EManual (Nandy et al., 2021), TechQA (Castelli et al., 2020), FinQA (Chen et al., 2021), TAT-QA (Zhu et al., 2021), ExpertQA (Malaviya et al., 2024), HAGRID (Kamalloo et al., 2023)), as well as one that was specifically adapted for RAG (DelucionQA (Sadat et al., 2023)). This measures the ability of a LLM to incorporate retrieved documents along with user queries to generate accurate answers for problems that require in-depth domain knowledge. As such, RAGbench is grouped by the needed domain knowledge: bio-medical research (PubmedQA, CovidQA), general knowledge (HotpotQA, MS Marco, HAGRID, ExperQA), legal contracts (CuAD), customer support (DelucionQA, EManual, TechQA), and finance (FinBench, TAT-QA). Finally, openhermes is a collection of GPT4 generated questions designed to emulate real user queries to an LLM.

A.2 SPROUT MODELS AND RESPONSE COLLECTION

Table 2 provides the models and their associated costs that a router trained on SPROUT can select between. The input and output token counts are collected by simply gathering the count of the tokenized queries and outputs of a model from its tokenizer. In order to emulate real-world use cases, responses from each LLM are collected using a corresponding *chat template* with a generic prompt and *zero shot prompting*.

Given the use of chat templates and zero-shot prompting, evaluation is challenging because model responses will not necessarily follow a specific format. To alleviate this, we adopt the evaluation

Table 1: Dataset Splits for SPROUT.

Benchmark	**Train**	**Validation**	**Test**
ragbench/expertqa	98	17	16
MATH (test)	1725	363	384
ragbench (emanual)	82	27	23
ragbench (cuad)	151	35	29
MuSR	178	35	35
MATH	5217	1061	1134
MuSR (team allocation)	157	52	41
ragbench (hagrid)	92	23	17
gpqa (extended)	368	89	84
MuSR (object placements)	169	47	34
ragbench (pubmedqa)	92	14	26
ragbench (hotpotqa)	89	22	21
ragbench (msmarco)	85	24	23
ragbench (techqa)	85	24	23
MMLU-Pro	8204	1784	1798
openhermes	13703	2917	2835
ragbench (tatqa)	90	17	25
ragbench (finqa)	97	15	20
ragbench (covidqa)	162	38	41
ragbench (delucionqa)	124	32	28
TOTAL	30968	6636	6637

Table 2: Models in SPROUT dataset and their API prices according to token counts.

Model	Input Token Cost (in $ per 1M tokens)	Output Token Cost (in $ per 1M tokens)
claude-3-5-sonnet-v1	3	15
titan-text-premier-v1	0.5	1.5
openai-gpt-4o	2.5	10
openai-gpt-4o-mini	0.15	0.6
granite-3-2b-instruct	0.1	0.1
granite-3-8b-instruct	0.2	0.2
llama-3-1-70b-instruct	0.9	0.9
llama-3-1-8b-instruct	0.2	0.2
llama-3-2-1b-instruct	0.06	0.06
llama-3-2-3b-instruct	0.06	0.06
llama-3-3-70b-instruct	0.9	0.9
mixtral-8x7b-instruct	0.6	0.6
llama-3-405b-instruct	3.5	3.5

protocol from MixEval (Ni et al., 2024) and use LLama-3.1-70B as a grader to score model queries against a given gold standard answer. The prompt format that we use is provided in D. Note that this prompt format needs to be converted to openai-api compatible messages while prompting the LLMs, which can be inferred from the special delimiters contained within the prompt format.

B Additional Plots and Experimental details

B.1 RouteBench

Figure 4 lays out the models and benchmarks present in the Routerbench dataset. To implement the transformer-based plug-in estimate of cost and accuracy, we utilize the `roberta-base` architecture with a learning rate of `3e-5` and a weight decay of 0.01. A training, validation, test split of 0.72, 0.8, 0.2 is used. Learning proceeds for 5 epochs, and the model with the best validation performance is saved at the end. To fit the KNN-based router, the OpenAI text-embedding-small-3 model is used, while the KNN regressor utilizes the 40-nearest neighbors measured by the 'cosine' similarity metric.

The same `roberta-base` parameters are used to fit the Roberta technique from RouteLLM (Ong et al., 2024). The matrix factorization method assumes that

$$\mathbf{P}(\text{GPT-4 Win}|q) = \sigma(w_2^T(v_{\text{GPT-4}} \odot (W_1^T v_q + b) - v_{\text{mixtral}} \odot (W_1^T v_q + b)))$$

where $v_{\text{GPT-4}}, v_{\text{mixtral}}$ are learnable embeddings of the model of interest. We use the `text-embeddder-small-3` from OpenAI to embed the queries, and a projection dimension of $d = 128$. The model is fit using Adam, with a learning rate of $3e-4$ and a weight decay of $1e-5$. To fit RoRF from not-diamond, we again use `text-embeddder-small-3` while the default parameters from Not-Diamond are used (max-depth = 20, 100 estimators).

Method	MMLU		MT-Bench		MBPP		HellaSwag		Winogrande		GSM8k		ARC	
	Perf↑	Cost↓	Perf↑	Cost↓	Perf↑	Cost↓	Perf↑	Cost↓	Perf↑	Cost↓	Perf↑	Cost↓	Perf↑	Cost↓
WizardLM 13B	0.568	0.122	0.796	0.006	0.364	0.011	0.636	0.727	0.512	0.040	0.510	0.354	0.660	0.068
Mistral 7B	0.562	0.081	0.779	0.003	0.349	0.006	0.541	0.485	0.562	0.027	0.409	0.210	0.642	0.046
Mixtral 8x7B	0.733	0.245	0.921	0.012	0.573	0.023	0.707	1.455	0.677	0.081	0.515	0.594	0.844	0.137
Code Llama 34B	0.569	0.317	0.796	0.015	0.465	0.021	0.525	1.882	0.617	0.104	0.462	0.752	0.644	0.177
Yi 34B	0.743	0.326	0.938	0.018	0.333	0.031	0.931	1.938	0.748	0.107	0.552	0.867	0.882	0.182
GPT-3.5	0.720	0.408	0.908	0.026	0.651	0.044	0.816	2.426	0.630	0.134	0.601	1.170	0.855	0.228
Claude Instant V1	0.384	0.327	0.863	0.030	0.550	0.064	0.801	1.943	0.512	0.108	0.626	1.300	0.821	0.183
Llama 70B	0.647	0.367	0.854	0.022	0.302	0.039	0.736	2.183	0.504	0.121	0.529	0.870	0.794	0.205
Claude V1	0.475	3.269	0.938	0.361	0.527	0.607	0.841	19.43	0.570	1.077	0.653	11.09	0.889	1.829
Claude V2	0.619	3.270	0.854	0.277	0.605	0.770	0.421	19.50	0.446	1.081	0.664	13.49	0.546	1.833
GPT-4	0.828	4.086	0.971	0.721	0.682	1.235	0.923	24.29	0.858	1.346	0.654	19.08	0.921	2.286
Oracle	**0.957**	0.297	**0.996**	0.052	**0.899**	0.041	**0.994**	0.860	**1.0**	0.042	**0.748**	1.282	**0.977**	0.091

Figure 4: Routerbench models and benchmarks (Hu et al. (2024) Table 1).

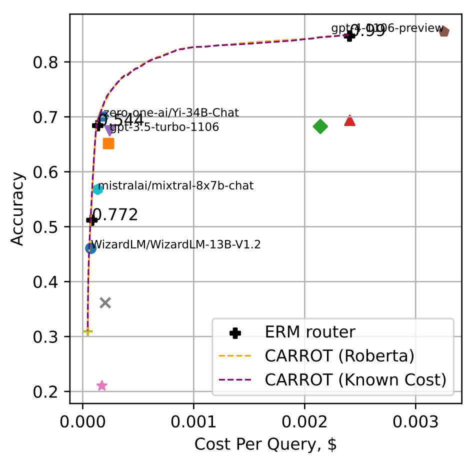

Figure 5: Router Bench Supplementary.

B.2 Open LLM Leaderboard V2

LLMs and costs: Table 3 gives all models used for the Open LLM Leaderboard experiment and their respective costs.

Table 3: Models used and their respective costs for the Open LLM Leaderboard experiment.

Model Name	Price (USD per 1M tokens)
NousResearch/Nous-Hermes-2-Mixtral-8x7B-DPO	0.6
01-ai/Yi-34B-Chat	0.8

Qwen/QwQ-32B-Preview	1.2
Qwen/Qwen2-72B-Instruct	0.9
Qwen/Qwen2.5-7B-Instruct	0.3
Qwen/Qwen2.5-72B-Instruct	1.2
alpindale/WizardLM-2-8x22B	1.2
deepseek-ai/deepseek-llm-67b-chat	0.9
google/gemma-2-27b-it	0.8
google/gemma-2-9b-it	0.3
google/gemma-2b-it	0.1
meta-llama/Llama-2-13b-chat-hf	0.3
meta-llama/Meta-Llama-3.1-70B-Instruct	0.9
mistralai/Mistral-7B-Instruct-v0.1	0.2
mistralai/Mistral-7B-Instruct-v0.2	0.2
mistralai/Mistral-7B-Instruct-v0.3	0.2
mistralai/Mixtral-8x7B-Instruct-v0.1	0.6
nvidia/Llama-3.1-Nemotron-70B-Instruct-HF	0.9

Model fitting: The model fitting details for baseline methods are all the same as in the RouterBench experiment (following the original implementations). To fit our methods, we employ some hyperparameter tuning for both KNN and `roberta-base`. For KNN, we employ 5-fold cross-validation using ROC-AUC and the possible number of neighbors as 2, 4, 8, 16, 32, 64, 128, 256, or 512. For `roberta-base` hyperparameter tuning, we train for 3k steps, using 20% of the training data for validation, a batch size of 8, and search for the best combination of learning rate, weight decay, and gradient accumulation steps in {5e-5, 1e-5}, {1e-2, 1e-4}, and {1, 2, 4, 8}. The final model is trained for 10k steps.

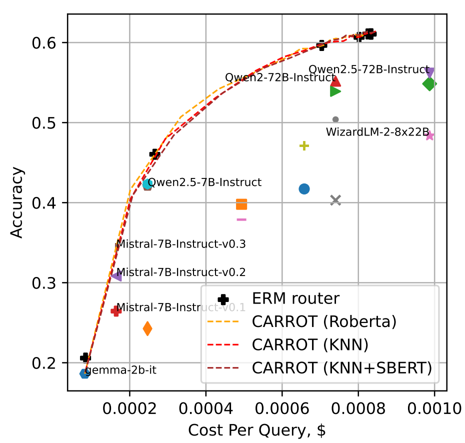

Figure 6: Open LLM leaderboard v2.

C SUPPLEMENTARY DEFINITIONS, RESULTS AND PROOFS

C.1 MINIMAX APPROACHES TO LEARNING THE RISK FUNCTIONS

In remark 3.10 we discussed the required condition for $\widehat{\Phi}$ so that the plug-in router has minimax rate optimal excess risk. In this section we show that estimating $\widehat{\Phi}$ using *local polynomial regression* (LPR) meets the requirement. To describe the LPR estimates consider a kernel $\psi : \mathbf{R}^d \to [0, \infty)$ that satisfies the regularity conditions described in the Definition C.2 in Appendix C with parameter $\max_k \gamma_k$ and define $\Theta(p)$ as the class of all p-degree polynomials from $\mathbf{R}^d$ to $\mathbf{R}$. For bandwidths

$h_k > 0; k \in [K_1]$ we define the LPR estimate as

$$[\widehat{\Phi}(x_0)]_{m,k} = \hat{\theta}_{x_0}^{(m,k)}(0);$$
$$\hat{\theta}_x^{(m,k)} \in \underset{\theta \in \Theta(p)}{\arg\min} \textstyle\sum_i \psi(\frac{X_i - x_0}{h})\big\{[Y_i]_{m,k} - \theta(X_i - x_0)\big\}^2. \tag{C.1}$$

In Theorem 3.2 of Audibert and Tsybakov (2007), a similar rate of convergence for LPR estimates is established. In their case, the losses were binary. For our instance, we assume that the Y_i are sub-Gaussian, but the conclusions are identical. We restate their result below.

Lemma C.1. *Assume that Y_i are sub-Gaussian random variables, i.e. there exist constants c_1 and c_2 such that*

$$P\big(\|Y_i\|_\infty > t \mid X\big) \le c_1 e^{-c_2 t^2}\,.$$

If ψ is regular (cf. Definition C.2) with parameter $\max_k \gamma_k$ and $p \ge \lfloor \max_k \gamma_k \rfloor$ then for $h_k = n^{-1/(2\gamma_k + d)}$ the Assumption 3.8 is satisfied with $a_{k,n} = n^{-\gamma_k/(2\gamma_k+d)}$, i.e. for some constants $\rho_1, \rho_2 > 0$ and any $n \ge 1$ and $t > 0$ and almost all X with respect to P_X we have the following concentration bound for $\widehat{\Phi}$:

$$\begin{aligned} \max_{P \in \mathcal{P}} P\Big\{ \max_{m,k} a_{k,n}^{-1} \big|[\widehat{\Phi}(X)]_{m,k} - [\Phi(X)]_{m,k}\big| \ge t \Big\} \\ \le \rho_1 \exp\big(-\rho_2 t^2\big)\,. \end{aligned} \tag{C.2}$$

This result is related to our Remark 3.10 about the rate-efficient estimation of routers. Estimating $\Phi(X)$ with an LPR and a suitable bandwidth and polynomial degree leads to our desired rate of convergence $a_{k,n} = n^{-\gamma_k/(2\gamma_k+d)}$ in Assumption 3.8.

C.2 Examples, additional assumptions and lemmas

Next, we describe the regularity conditions needed for local polynomial regression in eq. equation C.1 and equation C.1. These conditions are taken directly from Audibert and Tsybakov (2007, Section 3).

Definition C.2 (Kernel regularity). *For some $\beta > 0$ we say that a kernel $K : \mathbf{R}^d \to [0, \infty)$ satisfies the regularity condition with parameter β, or simply β-regular if the following are true:*

$$\begin{aligned} & \textit{for some } c > 0, K(x) \ge c, \ \textit{for } \|x\|_2 \le c\,, \\ & \textstyle\int K(x)dx = 1 \\ & \textstyle\int (1 + \|x\|_2^{4\beta}) K^2(x) dx < \infty, \\ & \sup_x (1 + \|x\|_2^{2\beta}) K(x) < \infty\,. \end{aligned}$$

An example of a kernel that satisfies these conditions is the Gaussian kernel: $K(x) = \prod_{j=1}^d \phi(x_j)$, where ϕ is the density of a standard normal distribution.

Next, we establish sufficient conditions for a class of distributions $\{p_\theta, \theta \in \mathbf{R}\}$ to satisfy the condition that $\mathrm{KL}(p_\theta, p_{\theta'}) \le K(\theta - \theta')^2$ for some $K > 0$ and any $\theta, \theta' \in \mathbf{R}$.

Lemma C.3. *Assume that a parametric family of distributions $\{p_\theta, \theta \in \mathbf{R}\}$ satisfies the following conditions:*

1. *The distributions have a density p_θ with respect to a base measure μ such that p_θ is twice continuously differentiable with respect to θ.*
2. $\int \partial_\theta p_\theta(x) d\mu(x) = \partial_\theta \int p_\theta(x) d\mu(x) = 0$
3. *For some $K > 0$ and all $\theta \in \mathbf{R}$ the $-\partial_\theta^2 \int \log p_\theta(x) p_\theta(x) d\mu(x) \le K$.*

Then $\mathrm{KL}(p_\theta, p_{\theta'}) \le \frac{K(\theta - \theta')^2}{2}$.

Some prominent examples of such family are location families of normal, binomial, Poisson distributions, etc.

Proof of the Lemma C.3. Notice that

$$
\begin{aligned}
&\mathrm{KL}(\mu_\theta, \mu_{\theta'})\\
&= \int p_\theta(x) \log \big\{ \tfrac{p_\theta(x)}{p_{\theta'}(x)} \big\} d\mu(x)\\
&= \int p_\theta(x) \big\{ \log p_\theta(x) - \log p_{\theta'}(x) \big\} d\mu(x)\\
&= \int p_\theta(x) \big\{ \log p_\theta(x) - \log p_\theta(x) - (\theta' - \theta)\partial_\theta \log p_\theta(x) - \tfrac{(\theta'-\theta)^2}{2} \partial_\theta^2 \log p_{\tilde\theta}(x) \big\} d\mu(x)
\end{aligned}
$$

Here, $\int p_\theta(x)\partial_\theta \log p_\theta(x) d\mu(x) = \int \partial_\theta p_\theta(x) d\mu(x) dx = 0$ and $-\int p_\theta(x) \partial_\theta^2 \log p_{\tilde\theta}(x) d\mu(x) \le K$. Thus, we have the upper bound $\mathrm{KL}(\mu_\theta, \mu_{\theta'}) \le \frac{K}{2}(\theta - \theta')^2$. □

C.3 PROOF OF LEMMA 2.1

Proof of Lemma 2.1. The μ-th risk

$$
\begin{aligned}
\mathcal{R}_P(g, \mu) &= \mathbf{E}\big[\mathbf{E}\big[Y\mu]_m \mid X\big] \mathbb{I}\{g(X) = m\}\big]\\
&= \mathbf{E}\big[\big\{ \textstyle\sum_{k=1}^K \mu_k [\Phi(X)]_{m,k} \big\} \mathbb{I}\{g(X) = m\}\big]
\end{aligned}
$$

is minimized at $g(X) = \arg\min_m \big\{ \sum_{k=1}^K \mu_k [\Phi(X)]_{m,k} \big\}$. □

C.4 THE UPPER BOUND

Lemma C.4. *Suppose that we have a function $f : \mathcal{X} \to \mathbf{R}^M$ for which we define the coordinate minimizer $g : \mathcal{X} \to [M]$ as $g(x) = \arg\min_m f_m(x)$ and the margin function*

$$
\Delta(x) = \begin{cases} \min_{m \neq g(x)} f_m(x) - f_{g(x)}(x) & \text{if } g(x) \neq [M] \\ 0 & \text{otherwise} . \end{cases}
$$

Assume that the margin condition is satisfied, i.e. there exist α, K_α such that

$$
P_X\big\{0 < \Delta(X) \le t\big\} \le K_\alpha t^\alpha . \tag{C.3}
$$

Additionally, assume that there exists an estimator $\widehat{f}$ of the function f such that it satisfies a concentration bound: for some $\rho_1, \rho_2 > 0$ and any $n \ge 1$ and $t > 0$ and almost all x with respect to P_X we have the following concentration bound for $\widehat{\Phi}$:

$$
P_{\mathcal{D}_n}\big\{\|\widehat{f}(x) - f(x)\|_\infty \ge t\big\} \le \rho_1 \exp\big(-\rho_2 a_n^{-2} t^2\big) , \tag{C.4}
$$

where $\{a_n; n \ge 1\} \subset \mathbf{R}$ is a sequence that decreases to zero. Then for $\widehat{g}(x) = \arg\min_m \widehat{f}_m(x)$ there exists a $K > 0$ such that for any $n \ge 1$ we have the upper bound

$$
\mathbf{E}_{\mathcal{D}_n}\big[\mathbf{E}_P\big[f_{\widehat{g}(X)}(X) - f_{g(X)}(X)\big]\big] \le K a_n^{1+\alpha} . \tag{C.5}
$$

Proof. For an $x \in \mathcal{X}$ define $\delta_m(x) = f_m(x) - f_{g(x)}(x)$. Since $g(x) = \arg\min_m f_m(x)$ we have $\delta_m(x) \ge 0$ for all m, $\min_m \delta_m(x) = 0$. Furthermore, define $h(x) = \arg\min\{m \neq g(x) : f_m(x)\}$, *i.e.* the coordinate of $f(x)$ where the second minimum is achieved. Clearly, $\delta_{h(x)}(x) = \Delta(x)$. With these definitions, lets break down the excess risk as:

$$
\begin{aligned}
&\mathbf{E}_{\mathcal{D}_n}\big[\mathbf{E}_P\big[f_{\widehat{g}(X)}(X) - f_{g(X)}(X)\big]\big]\\
&= \mathbf{E}_{\mathcal{D}_n}\big[\mathbf{E}_P\big[\textstyle\sum_{m=1}^M \{f_m(X) - f_{g(X)}(X)\} \mathbb{I}\{\widehat{g}(X) = m\}\big]\big]\\
&= \mathbf{E}_{\mathcal{D}_n}\big[\mathbf{E}_P\big[\textstyle\sum_{m=1}^M \{f_m(X) - f_{g(X)}(X)\} \mathbb{I}\{\widehat{g}(X) = m\} \mathbb{I}\{\Delta(X) \le \tau\}\big]\big]\\
&\quad + \textstyle\sum_{i \ge 1} \mathbf{E}_{\mathcal{D}_n}\big[\mathbf{E}_P\big[\textstyle\sum_{m=1}^M \{f_m(X) - f_{g(X)}(X)\} \mathbb{I}\{\widehat{g}(X) = m\} \mathbb{I}\{\tau 2^{i-1} < \Delta(X) \le \tau 2^i\}\big]\big]
\end{aligned} \tag{C.6}
$$

where $\tau = 2\rho_2^{-1/2} a_n$. We deal with the summands one by one. First, if $\Delta(X) = 0$ then all the coordinates of $f(X)$ are identical, which further implies that $f_m(X) - f_{g(X)}(X) = 0$ for any m. Thus,

$$
\begin{aligned}
&\mathbf{E}_{\mathcal{D}_n}\big[\mathbf{E}_P\big[\textstyle\sum_{m=1}^M \{f_m(X) - f_{g(X)}(X)\} \mathbb{I}\{\widehat{g}(X) = m\} \mathbb{I}\{\Delta(X) \le \tau\}\big]\big]\\
&= \mathbf{E}_{\mathcal{D}_n}\big[\mathbf{E}_P\big[\textstyle\sum_{m=1}^M \{f_m(X) - f_{g(X)}(X)\} \mathbb{I}\{\widehat{g}(X) = m\} \mathbb{I}\{0 < \Delta(X) \le \tau\}\big]\big]
\end{aligned}
$$

If $m = g(X)$ then the summand is zero. For the other cases, $\widehat{g}(X) = m$ if $\widehat{f}(X)$ has the minimum value at the m-th coordinate. This further implies $\widehat{f}_m(X) \le \widehat{f}_{g(X)}(X)$. The only way this could happen if $|\widehat{f}_m(X) - f_m(X)| \ge {}^{\delta_m(X)}/_2$ or $|\widehat{f}_{g(X)}(X) - f_{g(X)}(X)| \ge {}^{\delta_m(X)}/_2$. Otherwise, if both are $|\widehat{f}_m(X) - f_m(X)| < {}^{\delta_m(X)}/_2$ and $|\widehat{f}_{g(X)}(X) - f_{g(X)}(X)| < {}^{\delta_m(X)}/_2$ this necessarily implies

$$\begin{aligned}\widehat{f}_{g(X)}(X) &< f_{g(X)}(X) + \tfrac{\delta_m(X)}{2}\\ &= f_m(X) - \delta_m(X) + \tfrac{\delta_m(X)}{2}\\ &= f_m(X) - \tfrac{\delta_m(X)}{2} < \widehat{f}_m(X)\,,\end{aligned}$$

which means for $\widehat{f}(X)$ the minimum is not achieved at the m-th coordinate. Now, $|\widehat{f}_m(X) - f_m(X)| \ge {}^{\delta_m(X)}/_2$ or $|\widehat{f}_{g(X)}(X) - f_{g(X)}(X)| \ge {}^{\delta_m(X)}/_2$ implies $\|\widehat{f}(X) - f(X)\|_\infty \ge {}^{\delta_m(X)}/_2$. With these observations we split the expectation as

$$\begin{aligned}&\mathbf{E}_{\mathcal{D}_n}\big[\mathbf{E}_P\big[\{f_m(X) - f_{g(X)}(X)\}\mathbb{I}\{\widehat{g}(X) = m\}\mathbb{I}\{0 < \Delta(X) \le \tau\}\big]\big]\\ &= \mathbf{E}_{\mathcal{D}_n}\big[\mathbf{E}_P\big[\{f_m(X) - f_{g(X)}(X)\}\mathbb{I}\{\widehat{g}(X) = m = g(X)\}\mathbb{I}\{0 < \Delta(X) \le \tau\}\big]\big]\\ &\quad + \mathbf{E}_{\mathcal{D}_n}\big[\mathbf{E}_P\big[\{f_m(X) - f_{g(X)}(X)\}\mathbb{I}\{\widehat{g}(X) = m \ne g(X)\}\mathbb{I}\{0 < \Delta(X) \le \tau\}\big]\big]\end{aligned}$$

The first part is zero, whereas the second part further simplifies as:

$$\begin{aligned}&\mathbf{E}_{\mathcal{D}_n}\big[\mathbf{E}_P\big[\{f_m(X) - f_{g(X)}(X)\}\mathbb{I}\{\widehat{g}(X) = m \ne g(X)\}\mathbb{I}\{0 < \Delta(X) \le \tau\}\big]\big]\\ &\le \mathbf{E}_{\mathcal{D}_n}\big[\mathbf{E}_P\big[\{f_m(X) - f_{g(X)}(X)\}\mathbb{I}\big\{\|\widehat{f}(X) - f(X)\|_\infty \ge \tfrac{\delta_m(X)}{2}\big\}\mathbb{I}\{0 < \Delta(X) \le \tau\}\big]\big]\\ &= \mathbf{E}_P\big[\{f_m(X) - f_{g(X)}(X)\}\mathbf{E}_{\mathcal{D}_n}\big[\mathbb{I}\big\{\|\widehat{f}(X) - f(X)\|_\infty \ge \tfrac{\delta_m(X)}{2}\big\}\big]\mathbb{I}\{0 < \Delta(X) \le \tau\}\big]\\ &= \mathbf{E}_P\big[\delta_m(X)P_{\mathcal{D}_n}\big\{\|\widehat{f}(X) - f(X)\|_\infty \ge \tfrac{\delta_m(X)}{2}\big\}\mathbb{I}\{0 < \Delta(X) \le \tau\}\big]\\ &\le \mathbf{E}_P\big[\delta_m(X)\rho_1 e^{-\frac{\rho_2 a_n^{-2}\delta_m^2(X)}{4}}\mathbb{I}\{0 < \Delta(X) \le \tau\}\big] = \mathbf{E}_P\big[\delta_m(X)\rho_1 e^{-\frac{\delta_m^2(X)}{\tau^2}}\mathbb{I}\{0 < \Delta(X) \le \tau\}\big]\end{aligned}$$

Notice that $\delta_m(X) \ge \Delta(X)$ whenever $\Delta(X) > 0$. Thus, we perform a maximization on $\delta_m(X)e^{-\frac{\delta_m^2(X)}{\tau^2}}$ on the feasible set $\delta_m(X) \ge \Delta(X)$. Here, we use the result:

$$\max_{x \ge y} xe^{-\frac{x^2}{\tau^2}} \le \begin{cases}\frac{\tau}{\sqrt{2e}} & \text{if } \frac{\tau}{\sqrt{2}} \ge y\\ ye^{-\frac{y^2}{\tau^2}} & \text{otherwise}\,,\end{cases} \tag{C.7}$$

where $x = \delta_m(X)$ and $y = \Delta(X)$. Since $\Delta(X) \le \tau$ we have $\delta_m(X)e^{-\frac{\delta_m^2(X)}{\tau^2}} \le \tau$ and thus

$$\mathbf{E}_P\big[\delta_m(X)\rho_1 e^{-\frac{\delta_m^2(X)}{\tau^2}}\mathbb{I}\{0 < \Delta(X) \le \tau\}\big] \le \rho_1\tau P\{0 < \Delta(X) \le \tau\} = \rho_1\tau^{1+\alpha}\,.$$

This finally results in

$$\mathbf{E}_{\mathcal{D}_n}\big[\mathbf{E}_P\big[\textstyle\sum_{m=1}^M\{f_m(X) - f_{g(X)}(X)\}\mathbb{I}\{\widehat{g}(X) = m\}\mathbb{I}\{\Delta(X) \le \tau\}\big]\big] \le M\rho_1\tau^{1+\alpha}\,,$$

which takes care of the first summand in eq. equation C.6. Now, for an $i \ge 1$, let us consider the summand

$$\mathbf{E}_{\mathcal{D}_n}\big[\mathbf{E}_P\big[\textstyle\sum_{m=1}^M\{f_m(X) - f_{g(X)}(X)\}\mathbb{I}\{\widehat{g}(X) = m\}\mathbb{I}\{\tau 2^{i-1} < \Delta(X) \le \tau 2^i\}\big]\big]$$

Again, on the event $m = g(X)$ the the summand is zero and on the other cases we have $\|\widehat{f}(X) - f(X)\|_\infty \ge {}^{\delta_m(X)}/_2$. Thus, we write

$$\begin{aligned}&\mathbf{E}_{\mathcal{D}_n}\big[\mathbf{E}_P\big[\textstyle\sum_{m=1}^M\{f_m(X) - f_{g(X)}(X)\}\mathbb{I}\{\widehat{g}(X) = m\}\mathbb{I}\{\tau 2^{i-1} < \Delta(X) \le \tau 2^i\}\big]\big]\\ &\le \textstyle\sum_{m=1}^M \mathbf{E}_{\mathcal{D}_n}\big[\mathbf{E}_P\big[\delta_m(X)\mathbb{I}\big\{\|\widehat{f}(X) - f(X)\|_\infty \ge \tfrac{\delta_m(X)}{2}\big\}\mathbb{I}\{\tau 2^{i-1} < \Delta(X) \le \tau 2^i\}\big]\big]\\ &\le \textstyle\sum_{m=1}^M \mathbf{E}_P\big[\delta_m(X)\rho_1 e^{-\frac{\delta_m^2(X)}{\tau^2}}\mathbb{I}\{\tau 2^{i-1} < \Delta(X) \le \tau 2^i\}\big]\end{aligned}$$

Because $\Delta(X) \ge \tau 2^{i-1} > {}^{\tau}/_{\sqrt{2}}$ we again use the inequality in eq. equation C.7 to obtain

$$\begin{aligned}&\textstyle\sum_{m=1}^M \mathbf{E}_P\big[\delta_m(X)\rho_1 e^{-\frac{\delta_m^2(X)}{\tau^2}}\mathbb{I}\{\tau 2^{i-1} < \Delta(X) \le \tau 2^i\}\big]\\ &\le \textstyle\sum_{m=1}^M \mathbf{E}_P\big[\Delta(X)\rho_1 e^{-\frac{\Delta^2(X)}{\tau^2}}\mathbb{I}\{\tau 2^{i-1} < \Delta(X) \le \tau 2^i\}\big]\\ &\le \textstyle\sum_{m=1}^M \tau 2^i\rho_1 e^{-\frac{\tau^2 2^{2i-2}}{\tau^2}} P\{\tau 2^{i-1} < \Delta(X) \le \tau 2^i\}\\ &\le M\tau 2^i\rho_1 e^{-\frac{\tau^2 2^{2i-2}}{\tau^2}} P\{0 < \Delta(X) \le \tau 2^i\} = M\rho_1\tau^{1+\alpha}2^{i(1+\alpha)}e^{-2^{2i-2}}\end{aligned}$$

Combining all the upper bounds in equation C.6 we finally obtain

$$\mathbf{E}_{\mathcal{D}_n}\big[\mathbf{E}_P\big[f_{\widehat{g}(X)}(X) - f_{g(X)}(X)\big]\big] \le M\rho_1\tau^{1+\alpha}\big\{1 + \textstyle\sum_{i\ge 1} 2^{i(1+\alpha)}e^{-2^{2i-2}}\big\} \tag{C.8}$$

As $\sum_{i\ge 1} 2^{i(1+\alpha)}e^{-2^{2i-2}}$ is finite we have the result. □

Proof of Theorem 3.9. The proof of the upper bound follows directly from the lemma C.4 once we establish that for $a_n = \sum_{k=1}^{K_1} \mu_k a_{k,n}$ the following concentration holds: for constants $\rho_1, \rho_2 > 0$ and any $n \ge 1$ and $t > 0$ and almost all X with respect to P_X we have

$$\max_{P\in\mathcal{P}} P\big\{\max_m \big|\widehat{\eta}_{\mu,m}(X) - \eta^\star_{\mu,m}(X)\big| \ge t\big\} \le \rho_1 \exp\big(-\rho_2 a_n^{-2} t^2\big)\,. \tag{C.9}$$

To this end, notice that

$$\begin{aligned} &\max_m \big|\widehat{\eta}_{\mu,m}(X) - \eta_{\mu,m}(X)\big| \\ &\le \textstyle\sum_{k=1}^{K} \mu_k \max_m \big|[\widehat{\Phi}(X)]_{m,k} - [\Phi(X)]_{m,k}\big| \\ &= \textstyle\sum_{k=1}^{K_1} \mu_k \max_m \big|[\widehat{\Phi}(X)]_{m,k} - [\Phi(X)]_{m,k}\big| \end{aligned}$$

where the last equality holds because $[\widehat{\Phi}(X)]_{m,k} = [\Phi(X)]_{m,k}$ for $k \ge K_1 + 1$. Following this inequality, we have that for any $P \in \mathcal{P}$

$$\begin{aligned} &P\big\{\max_m \big|\widehat{\eta}_{\mu,m}(X) - \eta_{\mu,m}(X)\big| \ge K_1 t\big\} \\ &\le \textstyle\sum_{k=1}^{K_1} P\big\{\max_m \big|[\widehat{\Phi}(X)]_{m,k} - [\Phi(X)]_{m,k}\big| \ge \frac{t}{\mu_k}\big\} \\ &\le \textstyle\sum_{k=1}^{K_1} \rho_{k,1}\exp\big(-\rho_{k,2}\mu_k^{-2} a_{k,n}^{-2} t^2\big) \\ &\le \rho_1 \exp\big(-\rho_2 K_1^2 \{\wedge_{k=1}^{K_1} \mu_k^{-1} a_{k,n}^{-1}\}^2 t^2\big) \end{aligned}$$

where $\rho_1 = \frac{\max_{k\le K_1}\rho_{k,1}}{K_1}$ and $\rho_2 = K_1^{-2} \times \{\wedge_{k\le K_1}\rho_{k,2}\}$. Note that

$$K_1\{\wedge_{k=1}^{K_1} \mu_k^{-1} a_{k,n}^{-1}\}^{-1} = K_1 \max_{k=1}^{K_1} \mu_k a_{l,n} \ge \textstyle\sum_{k\le K_1} \mu_k a_{k,n} = a_n\,.$$

Thus,

$$\begin{aligned} &P\big\{\max_m \big|\widehat{\eta}_{\mu,m}(X) - \eta_{\mu,m}(X)\big| \ge K_1 t\big\} \\ &\le \rho_1\exp\big(-\rho_2 K_1^2\{\wedge_{k=1}^{K_1}\mu_k^{-1}a_{k,n}^{-1}\}^2 t^2\big) \le \rho_1\exp\big(-\rho_2 a_n^2 t^2\big)\,. \end{aligned}$$

□

C.5 The lower bound

To begin, we discuss the high-level proof strategy that will achieve our lower bound. Ultimately, for every $k \le K_1$ we shall establish that for any $\epsilon_k \in [0, 1]$ and $n \ge 1$

$$\min_{A_n\in\mathcal{A}_n}\max_{P\in\mathcal{P}} \mathcal{E}_P(\mu, A_n) \ge c_k\big\{\mu_k n^{-\frac{\gamma_k}{2\gamma_k+d}}\big\}^{1+\alpha}\,, \tag{C.10}$$

for some constant $c_k > 0$. Then, defining $c = \min\{c_k : k \le K_1\}$ we have the lower bound

$$\begin{aligned} \min_{A_n\in\mathcal{A}_n}\max_{P\in\mathcal{P}} \mathcal{E}_P(\mu, A_n) &\ge \max_{k\le K_1} c_k\big\{\mu_k n^{-\frac{\gamma_k}{2\gamma_k+d}}\big\}^{1+\alpha} \\ &\ge \max_{k\le K_1} c\big\{\mu_k n^{-\frac{\gamma_k}{2\gamma_k+d}}\big\}^{1+\alpha} \\ &\ge c\Big\{\textstyle\sum_{k\le K_1}\frac{\mu_k n^{-\frac{\gamma_k}{2\gamma_k+d}}}{K}\Big\}^{1+\alpha} \\ &\ge cK^{-1-\alpha}\Big\{\textstyle\sum_{k\le K_1}\mu_k n^{-\frac{\gamma_k}{2\gamma_k+d}}\Big\}^{1+\alpha}\,, \end{aligned}$$

which would complete the proof.

It remains to establish equation C.10 for each $k \in [K_1]$. To obtain this, we construct a finite family of probability measures $\mathcal{M}_r \subset \mathcal{P}$ (indexed by $[r]$) and study $\max\limits_{P\in\mathcal{M}_r}$. The technical tool which allows this to be fruitful is a generalized version of Fano's lemma.

Lemma C.5 (Generalized Fano's lemma). *Let $r \geq 2$ be an integer and let $\mathcal{M}_r \subset \mathcal{P}$ contains r probability measures indexed by $\{1, \ldots, r\}$ such that for a pseudo-metric d (i.e. $d(\theta, \theta') = 0$ if and only if $\theta = \theta'$) any $j \neq j'$*

$$d\big(\theta(P_j), \theta(P_{j'})\big) \geq \alpha_r, \ \textit{ and } \ KL(P_j, P_{j'}) \leq \beta_r\,.$$

Then

$$\max_j \mathbf{E}_{P_j}\big[d(\theta(P_j), \widehat{\theta})\big] \geq \tfrac{\alpha_r}{2}\big(1 - \tfrac{\beta_r + \log 2}{\log r}\big)\,.$$

In our construction $\theta(P^\sigma) = g^\star_{\mu,\sigma}$ and $d\big(\theta(P^{\sigma_0}), \theta(P^{\sigma_1})\big) = \mathcal{E}_{P^{\sigma_0}}(g^\star_{\mu,\sigma_1}, \mu)$.

Next, we lay out the template for constructing the family $\mathcal{M}_r$. Fix a $k_0 \in [K_1]$ and define the following.

Definition C.6. *1. For an $h = L \times \mu_{k_0}^{\frac{1}{\gamma_{k_0}}} n^{-\frac{1}{2\gamma_{k_0}+d}}$ ($L > 0$ is a constant to be decided later) define $m = \lfloor h^{-1} \rfloor$.*

2. *Define $\mathcal{G} = \big[\{ih + \frac{h}{2} : i = 0, \ldots, m-1\}^d\big]$ as a uniform grid in $[0,1]^d$ of size m^d and $\mathcal{G}_\epsilon$ as an ϵ-net in ℓ_∞ metric, i.e. $\mathcal{G}_\epsilon = \cup_{x \in \mathcal{G}} \mathcal{B}(x, \epsilon, \ell_\infty)$, where $\mathcal{B}(x, \epsilon, \ell_\infty) = \{y \in \mathcal{X} : \|x - y\|_\infty \leq \epsilon\}$.*
3. *Define $P_X = \mathrm{Unif}(\mathcal{G}_\epsilon)$. For such a distribution, note that $\mathrm{vol}(\mathcal{G}_\epsilon) = (m\epsilon)^d \leq (h^{-1}\epsilon)^d$, which implies that for all $x \in \mathcal{G}_\epsilon$ we have $p_X(x) = (h\epsilon^{-1})^d$. Setting $\epsilon = p_0^{-1/d} h \wedge \frac{h}{3}$ we have $p_X(x) \geq p_0$ that satisfies the strong density assumption for P_X.*
4. *Fix an $m_0 \leq m^d$ and consider $\mathcal{G}_0 \subset \mathcal{G}$ such that $|\mathcal{G}_0| = m_0$ and define $\mathcal{G}_1 = \mathcal{G} \backslash \mathcal{G}_0$.*
5. *For a function $\sigma : \mathcal{G}_0 \to [M]$ define*

$$\Phi^\sigma_{m,k}(x) = \begin{cases} \frac{1 - K_{\gamma,k_0}\mu_{k_0}^{-1}\epsilon^{\gamma_{k_0}}\mathbb{I}\{\sigma(y) = m\}}{2} & \textit{when } k = k_0,\ x \in \mathcal{B}(x, \epsilon, \ell_\infty) \textit{ for some } y \in \mathcal{G}_0, \\ \frac{1}{2} & \textit{elsewhere.} \end{cases} \tag{C.11}$$

6. *Consider a class of probability distributions $\{\mu_\theta : \theta \in \mathbf{R}\}$ defined on the same support range(ℓ) that have mean θ and satisfy $KL(\mu_\theta, \mu_{\theta'}) \leq c(\theta - \theta')^2$ for some $c > 0$. A sufficient condition for constricting such a family of distributions can be found in Lemma C.3. Some prominent examples of such family are location families of normal, binomial, Poisson distributions, etc. Define the probability $P^\sigma([Y]_{m,k} \mid X = x) \sim \mu_{\Phi^{(\sigma)}_{m,k}(x)}$.*

The following two lemmas (along with the observation on the strong density condition) will establish that for a given σ, the distribution over $\mathcal{X}, \mathcal{Y}$ given by $P^\sigma([Y]_{m,k} \mid X = x) \times \mathrm{Unif}[\mathcal{G}_\epsilon]$ is indeed a member of the class $\mathcal{P}$.

Lemma C.7. *Fix a choice for σ and let $\eta^\sigma_{\mu,m} = \sum_k \mu_k \Phi^\sigma_{k,m}(x)$, then $\eta^\sigma_{\mu,m}$ satisfies α-margin condition.*

Proof. To see that $\eta^\sigma_{\mu,m}$ satisfies α-margin condition, notice that

$$\eta^\sigma_{\mu,m}(x) = \begin{cases} \frac{1 - K_{\gamma,k_0}\epsilon^{\gamma_{k_0}}\mathbb{I}\{\sigma(y) = m\}}{2} & \text{when } x \in \mathcal{B}(x, \epsilon, \ell_\infty) \text{ for some } y \in \mathcal{G}_0, \\ \frac{1}{2} & \text{elsewhere.} \end{cases}$$

Thus, for every $x \in \mathcal{B}(y, \epsilon, \ell_\infty), y \in \mathcal{G}_0$ the $\Phi^\sigma_{\mu,m}(x) = \frac{1}{2}$ for all but one m and at $m = \sigma(x)$ the $\Phi^\sigma_{\mu,m}(x) = \frac{1 - K_{\gamma,k_0}\epsilon^{\gamma_{k_0}}}{2}$, leading to $\Delta^\sigma_\mu(x) = \frac{K_{\gamma,k_0}\epsilon^{\gamma_{k_0}}}{2}$ at those x, and at all other x we have $\Delta^\sigma_\mu(x) = 0$. This further implies $P_X(0 < \Delta^\sigma_\mu(X) \leq t) = 0$ whenever $t < \frac{K_{\gamma,k_0}\epsilon^{\gamma_{k_0}}}{2}$ and for $t \geq \frac{K_{\gamma,k_0}\epsilon^{\gamma_{k_0}}}{2}$ we have

$$\begin{aligned} P_X(0 < \Delta^\sigma(X) \leq t) &= P_X\big(\Phi^\sigma_m(X) \neq \tfrac{1}{2} \text{ for some } m \in [M]\big) \\ &\leq m_0\epsilon^d \leq K_\alpha\big(\tfrac{K_{\gamma,k_0}\epsilon^{\gamma_{k_0}}}{2}\big)^\alpha \end{aligned}$$

whenever

$$m_0 \leq K_\alpha 2^{-\alpha} K^\alpha_{\gamma,k_0}\epsilon^{\alpha\gamma_{k_0} - d}$$

We set $m_0 = \lfloor K_\alpha 2^{-\alpha} K^\alpha_{\gamma,k_0}\epsilon^{\alpha\gamma_{k_0} - d}\rfloor$ to meet the requirement. Since $d > \min_k \alpha\gamma_k$, for sufficiently small ϵ we have $m_0 \geq 8$. $\square$

Lemma C.8. *On the support of P_X the $\Phi^{\sigma}_{m,k}$ are $(\gamma_k, K_{\gamma,k})$ Hölder smooth.*

Proof. Note that the only way $\Phi^{\sigma}_{m,k}(x)$ and $\Phi^{\sigma}_{m,k}(x')$ can be different if $\|x - x'\|_\infty \geq \frac{h}{3}$. Since $\epsilon \leq \frac{h}{3}$ for such a choice, we have

$$\begin{aligned}|\Phi^{\sigma}_{m,k}(x) - \Phi^{\sigma}_{m,k}(x')| &\leq \tfrac{1}{2} K_{\gamma,k}\epsilon^{\beta} \\ &\leq K_{\gamma,k}(\tfrac{h}{3})^{\beta} \\ &\leq K_{\gamma,k}\|x - x'\|_\infty^{\beta} \leq K_{\gamma,k}\|x - x'\|_2^{\beta}\,.\end{aligned}$$

□

In order transfer the inequality in Fano's lemma to a statement on rate of convergence, we need an upper bound on $\mathrm{KL}(P^{\sigma_1}, P^{\sigma_2})$ and a lower bound on the semi-metric $\mathcal{E}_{P^{\sigma_0}}(\mu, g^\star_{\mu,\sigma_1})$. These are established in the next two lemmas.

Lemma C.9. *Consider the probability distribution P^{σ} for the random pair (X, Y) where $X \sim P_X$ and given X the $\{[Y]_{m,k}; m \in [M], k \leq K_1\}$ are all independent and distributed as $[Y]_{m,k} \mid X = x \sim \mu_{\Phi^{\sigma}_{m,k}(x)}$. Let C be a positive constant and $\delta(\sigma_1, \sigma_2) = \sum_{y \in \mathcal{G}_0} \mathbb{I}\{\sigma_1(y) \neq \sigma_2(y)\}$ the Hamming distance between σ_1 and σ_2. Then following upper bound holds on KL$(P^{\sigma_1}, P^{\sigma_2})$.*

$$\mathit{KL}(P^{\sigma_1}, P^{\sigma_2}) \leq C\mu_{k_0}^{-2} h^{2\gamma_{k_0}+d}\delta(\sigma_1, \sigma_2)$$

Proof.

$$\begin{aligned}&\mathrm{KL}(P^{\sigma_1}, P^{\sigma_2}) \\ &= \textstyle\int dP_X(x) \sum_{m=1}^M \sum_{k=1}^K \mathrm{KL}\big(\mu_{\Phi^{(\sigma_1)}_{m,k}(x)}, \mu_{\Phi^{(\sigma_2)}_{m,k}(x)}\big) \\ &\leq \textstyle\int dP_X(x) \sum_{m=1}^M \sum_{k=1}^K c\big(\Phi^{(\sigma_1)}_{m,k}(x) - \Phi^{(\sigma_2)}_{m,k}(x)\big)^2 \qquad (\mathrm{KL}(\mu_\theta, \mu_{\theta'}) \leq c(\theta - \theta')^2) \\ &= \textstyle\sum_{y \in \mathcal{G}_0} \epsilon^d \sum_{m=1}^M \frac{cK^2_{\gamma,k_0}\epsilon^{2\gamma_{k_0}}\mu_{k_0}^{-2}}{4}\big(\mathbb{I}\{\sigma_1(y) = m\} - \mathbb{I}\{\sigma_2(y) = m\}\big)^2 \\ &\leq \textstyle\frac{cK^2_{\gamma,k_0}}{4} \sum_{y \in \mathcal{G}_0} \mu_{k_0}^{-2}\epsilon^{2\gamma_{k_0}+d} \times \mathbb{I}\{\sigma_1(y) \neq \sigma_2(y)\} \\ &\leq C\mu_{k_0}^{-2} h^{2\gamma_{k_0}+d}\delta(\sigma_1, \sigma_2) \qquad (\text{because } \epsilon \leq \tfrac{h}{3})\end{aligned}$$

for some $C > 0$, where $\delta(\sigma_1, \sigma_2) = \sum_{y \in \mathcal{G}_0} \mathbb{I}\{\sigma_1(y) \neq \sigma_2(y)\}$ is the Hamming distances between σ_1 and σ_2. □

Now, we establish a closed form for the excess risk

$$\mathcal{E}_{P^{\sigma_0}}(\mu, g^\star_{\mu,\sigma_1}) = \mathbf{E}_{P^{\sigma_0}}(\mu, g^\star_{\mu,\sigma_1}) - \mathbf{E}_{P^{\sigma_0}}(\mu, g^\star_{\mu,\sigma_0})$$

where $g^\star_{\mu,\sigma_0}$ is the Bayes classifier for P^{σ_0} defined as $g^\star_{\mu,\sigma_0}(x) = \arg\min_m \Phi^{\sigma_0}_{\mu,m}(x)$.

Lemma C.10. *Let $\delta(\sigma_0, \sigma_1)$ denote the Hamming distance between σ_0 and σ_1 as before. Then*

$$\mathcal{E}_{P^{\sigma_0}}(\mu, g^\star_{\mu,\sigma_1}) = \frac{K_{\gamma,k_0}\epsilon^{\gamma_{k_0}+d}\delta(\sigma_0,\sigma_1)}{2}$$

Proof. For the purpose, notice that

$$g^\star_{\mu,\sigma}(x) = \sigma(y) \text{ whenever } x \in \mathcal{B}(x, \epsilon, \ell_\infty) \text{ for some } y \in \mathcal{G}_0\,.$$

This further implies

$$\begin{aligned}&\mathbf{E}_{P^{\sigma_0}}(\mu, g^\star_{\mu,\sigma_1}) \\ &= \textstyle\int dP_X(x) \sum_{m=1}^M \mathbb{I}\{g^\star_{\mu,\sigma_1}(x) = m\}\Phi^{\sigma_0}_{\mu,m}(x) \\ &= \textstyle\sum_{y \in \mathcal{G}_0} \epsilon^d \sum_{m=1}^M \mathbb{I}\{\sigma_1(y) = m\}\mu_{k_0}\frac{1}{2}\big\{1 - K_{\gamma,k_0}\mu_{k_0}^{-1}\epsilon^{\gamma_{k_0}}\mathbb{I}\{\sigma_0(y) = m\}\big\} \\ &\quad + \textstyle\sum_{y \in \mathcal{G}_0} \epsilon^d \sum_{m=1}^M \mathbb{I}\{\sigma_1(y) = m\} \sum_{k \neq k_0} \frac{\mu_k}{2} + \sum_{y \in \mathcal{G}_1} \epsilon^d \sum_{m=1}^M \mathbb{I}\{\sigma_1(y) = m\}\frac{1}{2} \\ &= -\textstyle\sum_{y \in \mathcal{G}_0} \sum_{m=1}^M \frac{K_{\gamma,k_0}\epsilon^{\gamma_{k_0}+d}}{2}\mathbb{I}\{\sigma_0(y) = \sigma_1(y) = m\} \\ &\quad + \textstyle\sum_{y \in \mathcal{G}_0 \cup \mathcal{G}_1} \epsilon^d \sum_{m=1}^M \mathbb{I}\{\sigma_1(y) = m\}\frac{1}{2} \\ &= -\textstyle\sum_{y \in \mathcal{G}_0} \sum_{m=1}^M \frac{K_{\gamma,k_0}\epsilon^{\gamma_{k_0}+d}}{2}\mathbb{I}\{\sigma_0(y) = \sigma_1(y) = m\} + \sum_{y \in \mathcal{G}_0 \cup \mathcal{G}_1} \frac{\epsilon^d}{2}\end{aligned}$$

By replacing σ_1 with σ_0 in the above calculations we obtain

$$\mathbf{E}_{P^{\sigma_0}}(\mu, g^\star_{\mu,\sigma_0}) = -\sum_{y\in\mathcal{G}_0}\sum_{m=1}^{M}\frac{K_{\gamma,k_0}\epsilon^{\gamma_{k_0}+d}}{2}\mathbb{I}\{\sigma_0(y)=m\} + \sum_{y\in\mathcal{G}_0\cup\mathcal{G}_1}\frac{\epsilon^d}{2}$$

and hence

$$\begin{aligned}
&\mathcal{E}_{P^{\sigma_0}}(g^\star_{\mu,\sigma_1},\mu)\\
&= \mathbf{E}_{P^{\sigma_0}}(g^\star_{\mu,\sigma_1},\mu) - \mathbf{E}_{P^{\sigma_0}}(g^\star_{\mu,\sigma_0},\mu)\\
&= \sum_{y\in\mathcal{G}_0}\sum_{m=1}^{M}\frac{K_{\gamma,k_0}\epsilon^{\gamma_{k_0}+d}}{2}\big\{\mathbb{I}\{\sigma_0(y)=m\} - \mathbb{I}\{\sigma_0(y)=\sigma_1(y)=m\}\big\}\\
&= \frac{K_{\gamma,k_0}\epsilon^{\gamma_{k_0}+d}}{2}\sum_{y\in\mathcal{G}_0}\sum_{m=1}^{M}\mathbb{I}\{\sigma_0(y)=m\}\times\mathbb{I}\{\sigma_1(y)\neq m\}\\
&= \frac{K_{\gamma,k_0}\epsilon^{\gamma_{k_0}+d}}{2}\sum_{y\in\mathcal{G}_0}\mathbb{I}\{\sigma_0(y)\neq\sigma_1(y)\}\\
&= \frac{K_{\gamma,k_0}\epsilon^{\gamma_{k_0}+d}\delta(\sigma_0,\sigma_1)}{2}\,.
\end{aligned}$$

□

The final technical ingredient we require is the Gilbert–Varshamov bound for linear codes.

Lemma C.11 (Gilbert–Varshamov bound)**.** *Consider the maximal $A_M(m_0,d)\subset[M]^{m_0}$ such that each element in C is at least d Hamming distance from each other, i.e. for any $\sigma_1,\sigma_2\in C$ we have $\delta(\sigma_1,\sigma_2)\geq d$. Then*

$$|A_M(m_0,d)| \geq \frac{M^{m_0}}{\sum_{i=0}^{d-1}\binom{m_0}{i}(M-1)^i}$$

Furthermore, when $M\geq 2$ and $0\leq p\leq 1-\frac{1}{M}$ we have $|A_M(m_0,pm_0)|\geq M^{m_0(1-h_M(p))}$ where $h_M(p) = \frac{p\log(M-1)-p\log p-(1-p)\log(1-p)}{\log M}$.

Proof of the Theorem 3.6. For the choice $p=\frac{1}{4}$ we have $-p\log p-(1-p)\log(1-p)\leq\frac{1}{4}$ and thus

$$h_M(p)\leq\frac{\log(M-1)}{4\log M}+\frac{1}{4\log M}\leq\frac{1}{4}+\frac{1}{4\log 2}\leq\frac{3}{4}\,.$$

Consequently, the lemma implies that we can find an $A_M(m_0,\frac{m_0}{4})\subset[M]^{m_0}$ such that $|A_M(m_0,\frac{m_0}{4})|\geq M^{\frac{m_0}{4}}$ whose each element is at least $\frac{m_0}{4}$ Hamming distance apart. For such a choice, define the collection of probabilities as $\mathcal{M}_r=\{P^\sigma:\sigma\in A_M(m_0,\frac{m_0}{4})\}$ leading to $r\geq M^{\frac{m_0}{4}}$. In the generalized Fano's lemma C.5 we require $r\geq 2$. To achieve that we simply set $m_0\geq 8$, as it implies $r\geq M^2\geq 4$.

Now we find lower bound α_r for the semi-metric and upper bound β_r for the Kulback-Leibler divergence. Let's start with the upper bound. Since $\mathrm{KL}(P^{\sigma_1},P^{\sigma_2})\leq C\mu_{k_0}^{-2}h^{2\gamma_{k_0}+d}\delta(\sigma_1,\sigma_2)$ for the joint distributions of the dataset $\mathcal{D}_n$ the Kulback-Leibler divergence between $\{P^{\sigma_1}\}^{\otimes n}$ and $\{P^{\sigma_2}\}^{\otimes n}$ is upper bounded as:

$$\begin{aligned}
&\mathrm{KL}\big(\{P^{\sigma_1}\}^{\otimes n},\{P^{\sigma_2}\}^{\otimes n}\big)\\
&= n\mathrm{KL}(P^{\sigma_1},P^{\sigma_2})\\
&\leq nC\mu_{k_0}^{-2}h^{2\gamma_{k_0}+d}\delta(\sigma_1,\sigma_2)\\
&= nC\mu_{k_0}^{-2}L^{2\gamma_{k_0}+d}\mu_{k_0}^{\frac{2\gamma_{k_0}+d}{\gamma_{k_0}}}n^{-\frac{2\gamma_{k_0}+d}{2\gamma_{k_0}+d}} \qquad (\text{because } h \text{ is defined as } L\times\mu_{k_0}^{\frac{1}{\gamma_{k_0}}}n^{-\frac{1}{2\gamma_{k_0}+d}})\\
&\leq CL^{2\gamma_{k_0}+d}\mu_{k_0}^{\frac{d}{\gamma_{k_0}}}\frac{\log r}{\log M} \qquad (\text{because } r\geq M^{\frac{m_0}{4}})\\
&\leq CL^{2\gamma_{k_0}+d}\frac{\log r}{\log M}=\beta_r
\end{aligned}$$

In the Lemma C.5 we would like $\frac{\beta_r+\log 2}{\log r}\leq\frac{3}{4}$ so that we have $1-\frac{\beta_r+\log 2}{\log r}\geq\frac{1}{4}$. Note that,

$$\begin{aligned}
\frac{\beta_r+\log 2}{\log r}-\frac{3}{4} &= \frac{\beta_r}{\log r}+\frac{\log 2}{\log r}-\frac{3}{4}\\
&= \frac{CL^{2\gamma_{k_0}+d}}{\log M}+\frac{\log 2}{\log 4}-\frac{3}{4} \qquad (\text{because } r\geq 4,\ \beta_r = CL^{2\gamma_{k_0}+d}\frac{\log r}{\log M})\\
&= \frac{CL^{2\gamma_{k_0}+d}}{\log M}-\frac{1}{4}\leq 0
\end{aligned}$$

for small $L > 0$. We set the L accordingly. Returning to the semi-metric, it is lower bounded as

$$
\begin{aligned}
d\big(\theta(P^{\sigma_0}), \theta(P^{\sigma_1})\big) &= \mathcal{E}_{P^{\sigma_0}}(g^{\star}_{\mu,\sigma_1}, \mu) \\
&\geq \tfrac{K_{\gamma,k_0}}{2}\epsilon^{\gamma_{k_0}+d}\delta(\sigma_0, \sigma_1) \\
&\geq \tfrac{K_{\gamma,k_0}}{2}\epsilon^{\gamma_{k_0}+d}\tfrac{m_0}{4} \\
&\geq \tfrac{K_{\gamma,k_0}}{8}\epsilon^{\gamma_{k_0}+d}K_\alpha 2^{-\alpha}K^{\alpha}_{\gamma,k_0}\epsilon^{\alpha\gamma_{k_0}-d} \\
&\quad \text{(because } m_0 = \lfloor K_\alpha 2^{-\alpha}K^{\alpha}_{\gamma,k_0}\epsilon^{\alpha\gamma_{k_0}-d}\rfloor) \\
&= c_1\epsilon^{(1+\alpha)\gamma_{k_0}} \\
&\geq c_2\big\{\mu_{k_0}n^{-\frac{\gamma_{k_0}}{2\gamma_{k_0}+d}}\big\}^{1+\alpha} = \alpha_r
\end{aligned}
$$

for some constants $c_1, c_2 > 0$. We plug in the lower and upper bound in Fano's lemma C.5 to obtain the lower bound:

$$
\tfrac{\alpha_r}{2}\big(1 - \tfrac{\beta_r + \log 2}{\log r}\big) \geq \frac{c_2\big\{\mu_{k_0}n^{-\frac{\gamma_{k_0}}{2\gamma_{k_0}+d}}\big\}^{1+\alpha}}{2} \times \tfrac{1}{4} \geq c_3\big\{\mu_{k_0}n^{-\frac{\gamma_{k_0}}{2\gamma_{k_0}+d}}\big\}^{1+\alpha}
$$

for some $c_3 > 0$ that is independent of both n and μ.

□

D JUDGE PROMPT

Prompt format for the Llama 3.1 70b evaluator instruct LLM

```
<dmf>user
I want you to act as a judge for how well a model did answering a user-defined task. You will be provided with a user-defined task that was given to the model, its golden answer(s), and the model's answer. The context of the task may not be given here. Your task is to judge how correct is the model's answer. Your task is to judge how correct the model's answer is based on the golden answer(s), without seeing the context of the task, and then give a correctness score. The correctness score should be one of the below numbers: 0.0 (totally wrong), 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, or 1.0 (totally right). You should also add a brief justification regarding how the model's answer conforms to or contradicts the golden answer(s).

Your response must follow the format
{{
   "correctness_score": your_score,
   "justification": your_justification
}}

Note that each one of the golden answers is considered correct. Thus if the Model's Answer matches anyone of the golden answers, it should be considered correct.
<dmf>assistant
Certainly! I will act as a judge and adhere to the format you requested. Under oath, I will judge without consulting my internal knowledge of facts under any circumstances.
<dmf>user
User-defined task --- Sandy bought 1 million Safe Moon tokens. She has 4 siblings. She wants to keep half of them to herself and divide the remaining tokens among her siblings. After splitting it up, how many more tokens will she have than any of her siblings?

Golden Answer(s) --- \<answer 1\> 375000

Model's Answer --- Sandy will have more tokens than any sibling by 3/8 million.
```

```
<dmf>assistant
{{
    "correctness_score": 1.0,
    "justification": "The golden answer states that Sandy will have 375,000 more tokens than any of her siblings, which is a precise numerical value. The model's answer translates this scenario into a fraction of the total, saying Sandy will have more tokens than any sibling by 3/8 million. 1 million tokens * 3/8 =375,000 tokens. So the model provided an answer in fractional form that, when converted to a numerical value, exactly matches the golden answer's quantity."
}}
<dmf>user
User-defined task --- Extract the names and emails mentioned in the following text. Richard Matthew Stallman, also known by his initials, rms,[1] is an American free software movement activist and programmer. He can be reached at rms@gnu.org based on the information on his webpage.

Golden Answer(s) --- {{"names":  ["Richard Matthew Stallman"], "emails": ["rms@gnu.org"]}}

Model's Answer --- names: Richard Matthew Stallman
<dmf>assistant
{{
    "correctness_score": 0.5,
    "justification": "The model correctly identified the names but failed to extract the emails."
}}
<dmf>user
User defined task --- Question: In 2015 Edgar Lungu became prime minister of?

Golden Answer:
<answer 1> Zambia; <answer 2> Zamibia; <answer 3> People of Zambia; <answer4> Zambian cuisine; <answer 5> Zambians; <answer 6> Culture of Zambia; <answer 7> Etymology of Zambia; <answer 8> Zambia; <answer 9> Health care in Zambia; <answer 10> ISO 3166-1:ZM; <answer11> Republic Of Zambia; <answer 12> Cuisine of Zambia; <answer 13> Sport in Zambia; <answer 14> Republic of Zambia; <answer 15> Zambian people ; <answer 16> Name of Zambia

Model's Answer: Prime Minister
<dmf>assistant
{{
    "correctness_score": 0.0,
    "justification": "The golden answers provide a detailed list of entities all relating to Zambia, indicating that Edgar Lungu became the leader (specifically, they mentioned \"prime minister\") of Zambia in 2015. The model's answer, \"Prime Minister,\" merely repeats part of the question without answering it."
}}
<dmf>user
User defined task --- Give a one-line concise summary of the following news
PARIS (Reuters) - French President Emmanuel Macron will set out plans for reforming the European Union on Tuesday, including proposals for a separate eurozone budget, despite a German election result that is likely to complicate his far-reaching ambitions. German Chancellor Angela Merkel s conservatives saw their support slide in Sunday s election, though they remain the biggest parliamentary bloc. She is expected to seek a coalition with the liberal Free Democrats (FDP) - who have criticized Macron s ideas for Europe - and the Greens. Elysee officials said Macron, who has promised sweeping reforms to Europe s monetary union in coordination with Merkel, hoped the issues to be raised in his speech would be taken into account in Germany s
```

```
    coalition negotiations. One Elysee official said a eurozone budget,
    one of Macron s most contentious ideas, would be necessary in due
    course and that the president would therefore raise the issue in his
    speech, to be delivered at the Sorbonne University in Paris. Since
    his election in May, Macron has made the overhaul of the EU and its
    institutions one of his major themes. As well as his eurozone budget
    idea, he wants to see the appointment of a eurozone finance minister
    and the creation of a rescue fund that would preemptively help
    countries facing economic trouble. Ahead of Sunday s election, Merkel
     had indicated her willingness to work with Macron on a reform agenda
    , even if her own ideas may not reach as far as his. But the election
     results have left Merkel facing a difficult coalition-building task
    which is in turn likely to limit her flexibility on Europe. A
    coalition of Merkel s CDU/CSU bloc, the FDP and the Greens is
    unprecedented at the national level - and any attempt by the
    chancellor and Macron to press for greater EU integration will face
    opposition from the new German lower house Bundestag. The FDP has
    called for a phasing out of Europe s ESM bailout fund and changes to
    EU treaties that would allow countries to leave the euro zone. And
    the far-right, eurosceptic Alternative for Germany is now the third
    biggest party in the Bundestag, further curbing Merkel s room for
    maneuver. But Elysee officials noted that the FDP had reaffirmed its
    attachment to the EU and to strong Franco-German relations, a point
    Macron was likely to emphasize in his speech while at the same time
    not seeking to impose anything on his partners. Macron, the sources
    said, would propose that the whole EU move forward together, and that
     those who did not want to should not stand in the way of those that
    did. Coming just two days after the German election, Macron s speech
    is likely to be interpreted in Germany as an attempt to shape the
    debate before the coalition talks begin in earnest. German coalition
    agreements are strict, with the contours set out in them limiting the
     government s room for maneuver.
In that respect, Merkel s ability to work with Macron on EU reform will
    be pre-determined by whatever coalition deal is struck. FDP leader
    Christian Lindner said on Monday he would not agree to any coalition
    that did not promise a change in the German government s direction.
    While that appeared to set the stage for tough talks, he also offered
     hope for Macron, saying when asked about the French president s
    eurozone budget ideas that the FDP had a strong interest in the
    strength of France.

Golden Answer(s) --- After German election, Macron to set out his vision
    for Europe

Model's Answer --- French President Emmanuel Macron to introduce plans
    for reforming the European Union amid the uncertain aftermath of
    German elections
<dmf>assistant
{{
    "correctness_score": 0.6,
    "justification": "While the golden answer is more concise, the model'
    s answer is largely similar to the golden answer in its meaning."
}}
<dmf>user
User defined task --- {MODELTASK}
Golden Answer(s) --- {GOLDENRESPONSE}
Model's Answer --- {MODELRESPONSE}
```