Likelihood Variance as Text Importance for Resampling Texts to Map Language Models

Momose Oyama[1,2] **Ryo Kishino**[1] **Hiroaki Yamagiwa**[1] **Hidetoshi Shimodaira**[1,2]
[1]Kyoto University [2]RIKEN
oyama.momose@sys.i.kyoto-u.ac.jp, kishino.ryo.32s@st.kyoto-u.ac.jp
{h.yamagiwa, shimo}@i.kyoto-u.ac.jp

Abstract

We address the computational cost of constructing a model map, which embeds diverse language models into a common space for comparison via KL divergence. The map relies on log-likelihoods over a large text set, making the cost proportional to the number of texts. To reduce this cost, we propose a resampling method that selects important texts with weights proportional to the variance of log-likelihoods across models for each text. Our method significantly reduces the number of required texts while preserving the accuracy of KL divergence estimates. Experiments show that it achieves comparable performance to uniform sampling with about half as many texts, and also facilitates efficient incorporation of new models into an existing map. These results enable scalable and efficient construction of language model maps.

1 Introduction

In recent years, an increasing number of studies have been conducted to systematically compare and organize language models (Yax et al., 2025; Zhuang et al., 2025; Zhu et al., 2025; Horwitz et al., 2025; Zhou et al., 2025; Harada et al., 2025; Pasquini et al., 2025). In this context, Oyama et al. (2025) proposed a method to estimate Kullback-Leibler (KL) divergence between language models based on log-likelihood vectors. In this method, language models with different architectures are embedded into a common space, and visualizing this creates a map of language models (Fig. 1).

The model map is constructed using a text set $D_N = \{x_1, \cdots, x_N\}$. Since D_N is a sample from a broader text population $D^\dagger = \{x_1^\dagger, \ldots, x_{N_0}^\dagger\}$, it introduces sampling error relative to the true relationships between models in the population[1]. Additionally, the computational cost, which is proportional to the number of texts N, is also an issue.

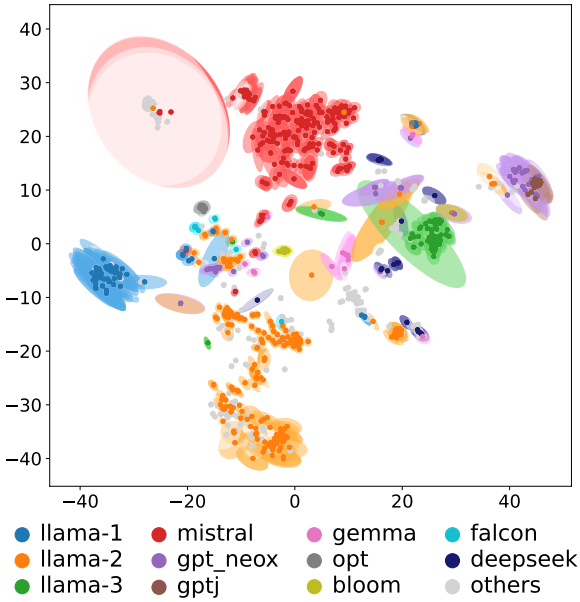


Figure 1: The model map calculated with D_N is visualized using t-SNE. Each point in the scatter plot corresponds to one of the 1,018 language models (Oyama et al., 2025) downloaded from Hugging Face. The sampling error for each point was estimated using the bootstrap resampling method and is shown as an ellipse. See Appendix B for details.

The primary motivation of this study is to reduce the computational costs associated with adding new models to the map. To address this issue, we propose a resampling procedure in which n texts are drawn with replacement from the dataset D_N. The resulting set, denoted by D_d^*, consists of d unique texts ($d \leq n$), which is then used to efficiently reconstruct the model map with the newly added models. When we discuss the distance between models calculated using log-likelihood vectors and the reliability of the model map, it is crucial to focus on the sampling error with respect to the population.

To estimate the resampling error of the distance

[1]When the entire Pile corpus is converted into 1,024-byte text chunks, the total number of texts is $N_0 = 5{,}703{,}791$.

Our code and data are available at https://github.com/shimo-lab/modelmap.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9453–9465

between models measured based on D_d^* relative to the true distances in the population, two types of errors need to be considered. The first is the **sampling error** that the original data D_N has with respect to the population $D^\dagger$, and the second is the **resampling error** incurred when selecting D_d^* from D_N.

Since uniform resampling from data D_N may waste many samples on less informative texts, we propose a method that preferentially resamples texts with large variance in log-likelihoods across models. This approach builds on the column sampling method of Drineas and Kannan (2001), originally proposed for approximating matrix products. Through experiments, we demonstrate that our text resampling method achieves an estimation error comparable to that of uniform random sampling while requiring only about half as many texts. Furthermore, it facilitates the efficient incorporation of new models into an existing model map while maintaining accuracy.

2 Background

2.1 Map of Language Models

The map of K language models $p_1, \ldots, p_K$ is constructed based on a text set $D_N = \{x_1, \cdots, x_N\}$. The log-likelihood of model p_i for text x_s is denoted by $\log p_i(x_s)$, and the log-likelihood matrix $\boldsymbol{L} \in \mathbb{R}^{K\times N}$ is composed of these elements. Let $\boldsymbol{Q} \in \mathbb{R}^{K\times N}$ be the matrix obtained by applying double centering[2] to matrix $\boldsymbol{L}$. The i-th row vector $\boldsymbol{q}_i \in \mathbb{R}^N$ of this matrix serves as the coordinate for language model p_i. Oyama et al. (2025) showed that the KL divergence between models p_i and p_j can be estimated by the following equation:

$$2\,\mathrm{KL}(p_i, p_j) \approx \|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2/N.$$

2.2 Length Squared Sampling

The idea of resampling text based on its importance is derived from column sampling methods, which probabilistically select a small number of columns from a matrix $\boldsymbol{A} = (A^{(1)}, \ldots, A^{(N)}) \in \mathbb{R}^{K\times N}$ to approximate the matrix product $\boldsymbol{A}\boldsymbol{A}^\top$.

In the representative Length Squared (LS) sampling method (Drineas and Kannan, 2001), each column $A^{(s)}$ is sampled with a probability proportional to the square of its Euclidean norm, $\|A^{(s)}\|^2$. This is known to minimize the expected Frobenius norm of the approximation error for $\boldsymbol{A}\boldsymbol{A}^\top$.

[2]Centering is performed row-wise (per model) and then column-wise (per text).

3 Resampling Texts for Model Map

3.1 Text Resampling Method

We apply the idea of LS sampling to reduce the number of texts used for the model map, while estimating the model distance $\|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2$ as accurately as possible. To determine the probability π_s that text x_s is resampled from dataset $D_N = \{x_1, \ldots, x_N\}$, we utilize the information in the double centered log-likelihood matrix $\boldsymbol{Q} \in \mathbb{R}^{K\times N}$.

LS Sampling. Following the column sampling framework of Drineas and Kannan (2001), we first propose to define the resampling probability for text x_s as

$$\pi_s \propto \|Q^{(s)}\|^2.$$

The squared norm $\|Q^{(s)}\|^2$ is proportional to the variance of the log-likelihoods for text x_s across models. This probability assignment is known to be optimal for approximating the inner product $\boldsymbol{q}_i^\top \boldsymbol{q}_j$, but not necessarily optimal for our goal of approximating the squared Euclidean distance $\|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2$.

KL Sampling. To directly address this goal, we introduce a novel resampling scheme, which we call KL sampling. Here, the resampling probability for text x_s is defined as

$$\pi_s \propto \sqrt{\sum_{i,j=1}^{K} (q_i(x_s) - q_j(x_s))^4}.$$

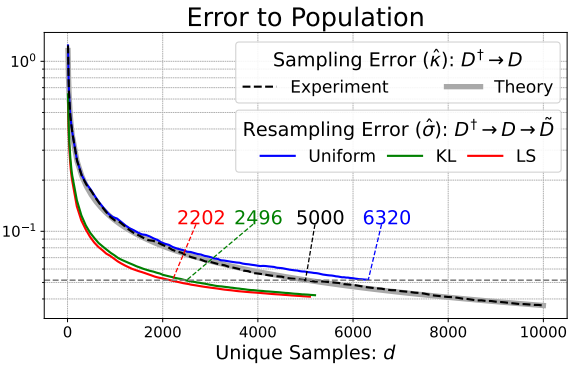
Here, $q_i(x_s)$ represents the (i, s)-th component of matrix $\boldsymbol{Q} \in \mathbb{R}^{K\times N}$. We show in Appendix C that this method yields an optimal approximation of the squared Euclidean distance $\|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2$.

Baseline: Uniform Sampling. All texts are resampled with equal probability $\pi_s = 1/N$.

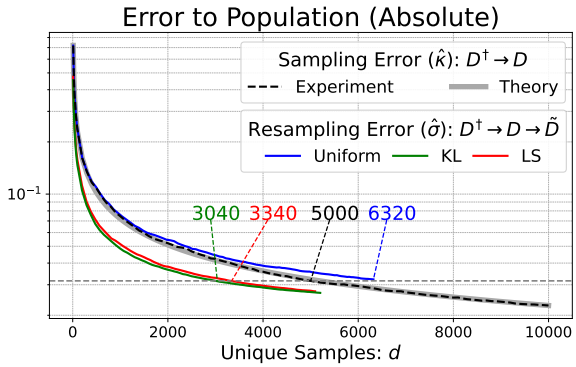
3.2 Model Map with Resampled Texts

Let $\widetilde{D}_n$ be the dataset obtained by resampling n texts from D_N, and the set of d unique texts in $\widetilde{D}_n$ be $D_d^* = \{x_{u_1}, \ldots, x_{u_d}\}$. We denote $c(u_t)$ as the number of times each text x_{u_t} was resampled, such that $\sum_{t=1}^{d} c(u_t) = n$.

We construct the log-likelihood matrix $\boldsymbol{L}_d \in \mathbb{R}^{K\times d}$ using the d resampled texts and obtain $\widetilde{\boldsymbol{Q}}_d$ by double centering. In row-wise centering, the column $L^{(u_t)}$ corresponding to the resampled text x_{u_t} should be weighted by $w_{u_t} = c(u_t)/n\pi_{u_t}$, based on the resampling probability π_s and the number of times it was resampled $c(u_t)$. Let $\widetilde{\boldsymbol{q}}_i \in \mathbb{R}^d$ be the row vector of $\widetilde{\boldsymbol{Q}}_d$. The distance between model



(a) Relative error $\widetilde{e}_{ij} = (\widetilde{g}_{ij} - g_{ij})/\max(g_{ij}, \varepsilon_0)$, normalized by g_{ij} for each model pair.



(b) Absolute error $\widetilde{e}_{ij} = (\widetilde{g}_{ij} - g_{ij})/C$ for each model pair, where $C = \sqrt{\frac{1}{K^2}\sum_{i,j=1}^{K} g_{ij}^2}$.

Figure 2: The error (RMSE) aggregated over all pairs of models, defined as $\hat{\sigma}_{\text{Method},n}$ for resampling methods and $\hat{\kappa}_d$ for ordinary sampling, plotted against the number of unique texts d. The errors are averaged over $R = 100$ trials. Panels show evaluation based on (a) relative error for each model pair (Section 4.1), and (b) absolute error (Section 4.2). In both evaluations, LS sampling and KL sampling show similar performance. This paper primarily explains the simpler LS sampling method.

d	500	2,500	5,000	7,500
Uniform	582 ± 4	$2{,}877 \pm 18$	N/A	N/A
KL	196 ± 2	$1{,}070 \pm 10$	$2{,}496 \pm 22$	$4{,}780 \pm 33$
LS	195 ± 2	895 ± 9	$2{,}202 \pm 19$	$4{,}229 \pm 30$

Table 1: Average and standard deviation of the number of unique texts d' required by each resampling method to achieve estimation accuracy comparable to that of ordinary sampling with d texts[3].

p_i and p_j is calculated as the weighted squared Euclidean distance using $\boldsymbol{w}_d = (w_{u_1}, \ldots, w_{u_d})$:

$$\|\widetilde{\boldsymbol{q}}_i - \widetilde{\boldsymbol{q}}_j\|^2_{\boldsymbol{w}_d} := \sum_{t=1}^{d} \frac{c(u_t)}{n\pi_{u_t}} \left(\widetilde{q}_i(x_{u_t}) - \widetilde{q}_j(x_{u_t})\right)^2,$$

where $\widetilde{q}_i(x_{u_t})$ is the (i,t)-th component of matrix $\widetilde{\boldsymbol{Q}}_d \in \mathbb{R}^{K \times d}$. This weighted distance serves as an estimate of the original distance defined under uniform sampling without weights.

4 Evaluation of Resampling Methods

4.1 Error of Resampled Estimates

The dataset D_N is a random sample from the text population $D^\dagger$, and $\widetilde{D}_n$ is resampled dataset from D_N. The squared Euclidean distances computed from $D^\dagger$, D_N, and $\widetilde{D}_n$ are, respectively, $g^\dagger_{ij} = (N/N_0)\|\boldsymbol{q}^\dagger_i - \boldsymbol{q}^\dagger_j\|^2$, $g_{ij} = \|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2$, $\widetilde{g}_{ij} = \|\widetilde{\boldsymbol{q}}_i - \widetilde{\boldsymbol{q}}_j\|^2_{\boldsymbol{w}_d}$, where $\boldsymbol{q}^\dagger_i \in \mathbb{R}^{N_0}$ denotes the coordinate of p_i in $D^\dagger$, and the scaling ensures comparability across different dimensionalities[4].

Our objective is to evaluate the error, namely how much $\widetilde{g}_{ij}$ deviates from the population value $g^\dagger_{ij}$, which cannot be directly computed in practice. Since $N_0 \gg N$, direct computation of $g^\dagger_{ij}$ is infeasible, and we therefore estimate the error using D_N. Details of this estimation procedure are provided in Appendix D.

Resampling Error to Population ($\hat{\sigma}_{\text{Method},n}$). Let $\hat{\sigma}_{\text{Method},n}$ (Method $\in \{\text{LS}, \text{KL}, \text{unif}\}$) be the estimated error relative to the true value $g^\dagger_{ij}$ in the population $D^\dagger$. This is estimated by considering the following two errors:

$$\hat{\sigma}_{\text{Method},n} = \sqrt{\tau^2_{\text{unif},N} + \tau^2_{\text{Method},n}}.$$

Here, $\tau_{\text{unif},N}$ is the bootstrap estimate (Efron and Tibshirani, 1994) of the sampling error that D_N itself has with respect to the population $D^\dagger$. Similarly, $\tau_{\text{Method},n}$ is the Root Mean Squared Error (RMSE) of the resampling with respect to D_N. It is calculated from the aggregated MSE of the relative error $\widetilde{e}_{ij} = (\widetilde{g}_{ij} - g_{ij})/\max(g_{ij}, \varepsilon_0)$ as

$$\tau^2_{\text{Method},n} = \frac{1}{K^2 R} \sum_{i,j=1}^{K} \sum_{r=1}^{R} \left(\widetilde{e}^{(r)}_{ij}\right)^2,$$

where $\widetilde{e}^{(r)}_{ij}$ denotes the relative error obtained in the r-th of R independent resampling trials. In the

[3]The values are computed from the resampled dataset $\widetilde{D}_{n'}$, where n' is the smallest resample size such that $\hat{\sigma}_{\text{Method},n'} \leq \hat{\kappa}_d$. Here, $\hat{\kappa}_d$ denotes the baseline error when d unique texts are directly sampled.

[4]$g^\dagger_{ij}$ and $\widetilde{g}_{ij}$ are rescaled to match the scale of g_{ij}.

experiments, we set $\varepsilon_0 = 10^{-3}$, $R = 100$, and varied n from 10 to 10,000.

Baseline: Sampling Error ($\hat{\kappa}_n$). Let κ_n be the aggregated relative error of g_{ij} with respect to the true value $g^{\dagger}_{ij}$ in the population. In this sampling, all n texts are unique, so the number of unique texts is $d = n$. The bootstrap estimate (Efron and Tibshirani, 1994) of κ_n is $\hat{\kappa}_n = \tau_{\text{unif},n}$.

Results. Figure 2a shows the relationship between the number of unique texts d and the estimated error with respect to the population for each resampling method. The dotted line represents the estimated error $\hat{\kappa}_d$ when d texts are directly randomly sampled from the population $D^{\dagger}$, and the thick gray solid line shows its theoretical value, computed as $\sqrt{N/d}\,\hat{\kappa}_N$. The colored solid lines represent the estimated error $\hat{\sigma}_{\text{Method},n}$ when resampling from D_N using each method. Comparing the blue solid line (Uniform resampling from D_N) and the black dotted line (direct sampling from $D^{\dagger}$), Uniform resampling from D_N results in an error comparable to directly sampling the same number of texts from the population[5]. In contrast, LS sampling and KL sampling achieve, with a smaller number of unique texts, an error comparable to the estimated error $\hat{\kappa}_d$ of random sampling d texts from the population. As can be seen from Table 1, an error comparable to $\hat{\kappa}_{5,000}$ is achieved with LS sampling using an average of $d = 2{,}202$ texts, and with KL sampling using an average of $d = 2{,}496$ texts, which is about half the number of texts required by uniform sampling from the population to achieve similar error. These results suggest that LS sampling and KL sampling can achieve comparable estimation accuracy with about half the number of unique texts by selecting important texts.

4.2 Discussion on LS and KL Sampling

In Section 4.1, the error was normalized by the magnitude of KL divergence for each model pair to evaluate the relative error with respect to KL divergence. In contrast, the KL sampling method directly optimizes $\mathbb{E}[\sum_{i,j=1}^{K}(\widetilde{g}_{ij} - g_{ij})^2 \mid D_N]$, which corresponds to minimizing the sum of absolute squared errors. Accordingly, we evaluate the error without normalization for each model pair. Specifically, we redefine the error as $\widetilde{e}_{ij} = (\widetilde{g}_{ij} - g_{ij})/C$, with $C = \sqrt{\frac{1}{K^2}\sum_{i,j=1}^{K} g_{ij}^2}$.

[5]In Uniform resampling, weighting according to the number of duplicates slightly degrades performance.

The results of this evaluation are shown in Fig. 2b. While KL sampling yields slightly smaller errors under the absolute-error evaluation (Fig. 2b), LS sampling can be slightly better under the relative-error evaluation (Fig. 2a). In both cases, however, the performance difference is very small. Therefore, we focus on the simpler LS sampling method, since it achieves performance comparable to KL sampling.

5 Efficiency of Resampled Text Subset

5.1 Resampling Error of Model Map

Settings. We resample n texts with replacement from D_N with $N = 10{,}000$, resulting in d unique texts[6]. We adopt LS sampling as the resampling method. Model coordinates are computed from log-likelihoods over the sampled texts and visualized using t-SNE. To evaluate variability due to resampling, we repeated the process 100 times per setting. Here, d denotes the average number of unique texts across trials. See Appendix B.2 for details.

Results. Figure 3(a) shows three model maps; ellipses indicate the standard deviation of model positions. The first two maps show similar stability, indicating that LS sampling achieves robustness comparable to that of uniform sampling while requiring fewer texts.

5.2 Adding New Models

We tested whether a small set of texts selected by LS sampling is sufficient for incorporating new models into the existing model map.

Settings. From the log-likelihoods of 898 models created before April 10, 2024, we sampled $n = 2{,}900$ texts via LS sampling, resulting in $d = 2{,}192$ unique texts. The remaining 120 models were treated as new additions[7]. This setting uses a single resampling trial, so d is the actual number of unique texts. We computed log-likelihoods for the newly added models using only these texts and visualized the updated map. To quantitatively evaluate efficiency, we compared the computation time of log-likelihoods using the 2,192 texts selected by LS sampling with that using all 10,000 texts, and also calculated the correlation between

[6]To verify that this set of 10,000 texts is not biased, we repeated the experiments with another disjoint set of 10,000 texts. See Appendix B.1 for details.

[7]Model creation dates were obtained using the Hugging Face API.

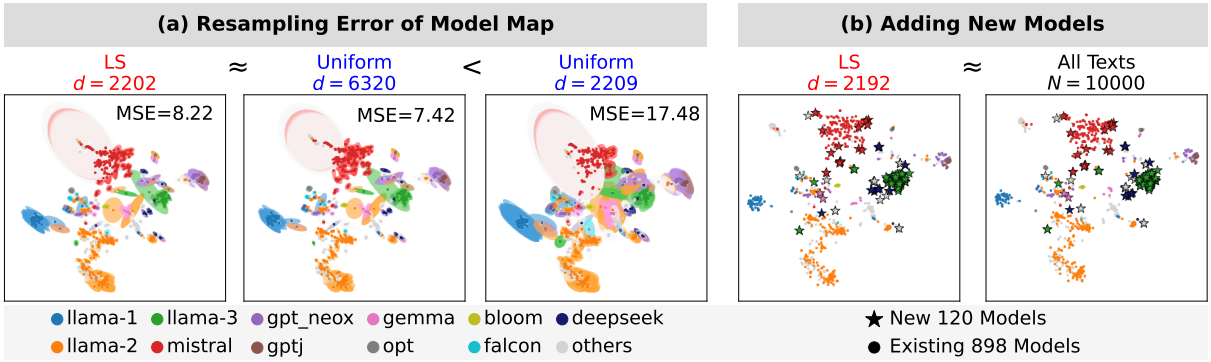


Figure 3: **(a)** Model maps based on LS ($n = 2{,}900, d = 2{,}202$), uniform ($n = 10{,}000, d = 6{,}320$), and uniform ($n = 2{,}500, d = 2{,}209$) sampling. Each map shows the mean coordinates and their variability as ellipses across 100 trials. The mean squared error (MSE) of each setting, summarized over the 100 trials, is also indicated in the plots for quantitative comparison. **(b)** Model maps with 120 new models added to existing 898 models. The left panel uses $d = 2{,}192$ unique texts selected by LS sampling with $n = 2{,}900$. The right panel uses all $N = 10{,}000$ texts.

New Models	LS $d = 2{,}192$	All $d = 10{,}000$	KL Corr.[8]
DeepSeek-R1-Distill-Llama-8B	1.94 min	8.55 min	0.996
Meta-Llama-3-8B	1.84 min	8.06 min	0.997
Phi-3-medium-128k-instruct	4.48 min	18.88 min	0.996
Mistral-7B-v0.3	2.11 min	8.98 min	0.997
Qwen2-1.5B	1.08 min	4.78 min	0.997

Table 2: Computation time of log-likelihoods using the 2,192-text subset selected by LS sampling and the full 10,000-text set. All computations were performed on a single RTX 6000 Ada GPU.

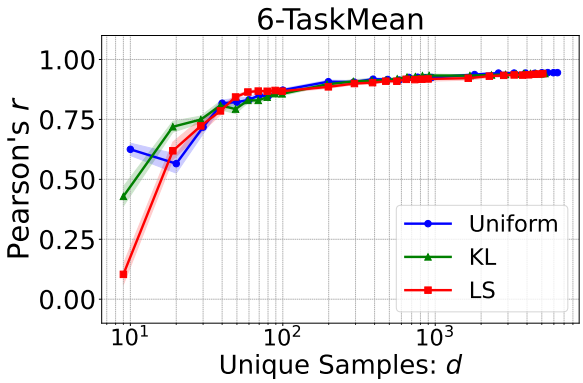


Figure 4: Pearson's correlation r between predicted and actual scores on the average of six downstream tasks, shown as a function of the number of unique texts d.

the estimated KL divergences. This evaluation was conducted on five representative models among the 120 new ones.

Results. As shown in Fig. 3(b), the result closely matches the full map based on all $N = 10{,}000$ texts, indicating that reliable placement is achievable with a small subset. Table 2 shows that, when using the 2,192 texts selected by LS sampling, log-likelihood computation is at least four times faster. Moreover, this efficiency comes with almost no loss in accuracy: the KL divergences estimated from this small subset are nearly identical to those from the full dataset, achieving a Pearson correlation of ~ 0.997.

5.3 Prediction of Downstream Performance

Following Oyama et al. (2025), we used model coordinates based on d unique texts to predict the average performance on six downstream tasks (Fig. 4). Although the resampling methods differ in KL estimation, prediction performance is nearly identical across methods, suggesting that the resulting coordinates span comparable subspaces. At $d = 1{,}000$, all methods achieve $r \approx 0.92$–0.93, indicating that even a relatively small number of texts suffices for accurate prediction. See Appendix E for details.

[8]Pearson correlation between KL divergences obtained with the 2,192-text subset and those with the full 10,000-text set, computed for each new model across its 898 divergences with the existing models.

6 Conclusion

We proposed text selection methods to reduce the computational cost of constructing model maps. Our experiments showed that these methods achieve estimation accuracy comparable to uniform sampling while requiring only about half as many texts, and that they also enable efficient incorporation of new models into an existing map. These findings facilitate more efficient comparative analysis of large-scale language models.

Limitations

- The aspect of data sampling has not been thoroughly explored in this study, apart from the simple experiment presented in Appendix B.1. Future work is needed to understand the extent to which our discussions generalize to datasets different from the D_N used here.

- Regarding the experiments for adding new models, we did not include a detailed discussion on the number or types of the new models added.

- While LS and KL sampling both outperformed uniform sampling in KL divergence estimation, the difference disappeared in downstream performance prediction, where all methods performed similarly. This suggests that improvements in distance estimation do not necessarily lead to gains in downstream utility, and the relationship between the two warrants further investigation.

Acknowledgments

This study was partially supported by JSPS KAKENHI 22H05106, 23H03355, 25K24366, JST CREST JPMJCR21N3, JST BOOST JPMJBS2407.

References

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open LLM Leaderboard. https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard.

Peter J Bickel and Anat Sakov. 2008. On the Choice of m in the m out of n Bootstrap and Confidence Bounds for Extrema. *Statistica Sinica*, pages 967–985.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *Preprint*, arXiv:1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *Preprint*, arXiv:2110.14168.

Petros Drineas and Ravi Kannan. 2001. Fast Monte-Carlo Algorithms for Approximate Matrix Multiplication. In *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science*, pages 452–459.

B. Efron and R.J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *Preprint*, arXiv:2101.00027.

Yuto Harada, Yusuke Yamauchi, Yusuke Oda, Yohei Oseki, Yusuke Miyao, and Yu Takagi. 2025. Massive Supervised Fine-tuning Experiments Reveal How Data, Layer, and Training Factors Shape LLM Alignment Quality. *Preprint*, arXiv:2506.14681.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *Preprint*, arXiv:2009.03300.

Eliahu Horwitz, Nitzan Kurer, Jonathan Kahana, Liel Amar, and Yedid Hoshen. 2025. We should chart an atlas of all the world's models. *Preprint*, arXiv:2503.10633.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *Preprint*, arXiv:2109.07958.

Momose Oyama, Hiroaki Yamagiwa, Yusuke Takase, and Hidetoshi Shimodaira. 2025. Mapping 1,000+ Language Models via the Log-Likelihood Vector. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32983–33038.

Dario Pasquini, Evgenios M. Kornaropoulos, and Giuseppe Ateniese. 2025. LLMmap: Fingerprinting for Large Language Models. In *Proceedings of the 34th USENIX Security Symposium*, pages 299–318.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *Preprint*, arXiv:1907.10641.

Peter H. Schönemann. 1966. A Generalized Solution of the Orthogonal Procrustes Problem. *Psychometrika*, 31:1–10.

Hidetoshi Shimodaira. 2014. Higher-Order Accuracy of Multiscale-Double Bootstrap for Testing Regions. *Journal of Multivariate Analysis*, 130:208–223.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Gaël Varoquaux, Lars Buitinck, Gilles Louppe, Olivier Grisel, Fabian Pedregosa, and Andreas Mueller. 2015. Scikit-learn: Machine Learning Without Learning the Machinery. *GetMobile Mob. Comput. Commun.*, 19(1):29–33.

Nicolas Yax, Pierre-Yves Oudeyer, and Stefano Palminteri. 2025. PhyloLM: Inferring the phylogeny of large language models and predicting their performances in benchmarks. In *Proceedings of the 13th International Conference on Learning Representations*.

Robert S. Yuill. 1971. The Standard Deviational Ellipse; An Updated Tool for Spatial Description. *Geografiska Annaler: Series B, Human Geography*, 53(1):28–39.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? *Preprint*, arXiv:1905.07830.

Xinyu Zhou, Delong Chen, Samuel Cahyawijaya, Xufeng Duan, and Zhenguang Cai. 2025. Linguistic minimal pairs elicit linguistic similarity in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6866–6888.

Sally Zhu, Ahmed M Ahmed, Rohith Kuditipudi, and Percy Liang. 2025. Independence tests for language models. In *Proceedings of the 42nd International Conference on Machine Learning*.

Richard Zhuang, Tianhao Wu, Zhaojin Wen, Andrew Li, Jiantao Jiao, and Kannan Ramchandran. 2025. EmbedLLM: Learning compact representations of large language models. In *Proceedings of the 13th International Conference on Learning Representations*.

A Author Contributions

M.O. conceived the overall research idea of reducing computational cost via resampling, developed the LS sampling algorithm using matrix approximation techniques, and implemented all resampling methods and experiments related to model distance estimation and map construction. R.K. proposed the KL sampling method and proved its optimality. H.S. developed the statistical framework for analyzing sampling and resampling errors, and provided theoretical justification. H.Y. conducted the experiments on downstream task performance prediction. All authors contributed to writing the manuscript. M.O. coordinated the writing process and took the lead on the overall composition, while each author wrote the parts corresponding to their contributions. The project was supervised by M.O. and H.S.

B Details of Model Map Construction

B.1 Log-Likelihood Data of Language Models

In our experiment, we used log-likelihood data (Oyama et al., 2025)[9] for $K = 1{,}018$ language models, calculated on $N = 10{,}000$ texts extracted from the Pile (Gao et al., 2020).

In the experiment described in Section 5.1 for adding new models to the model map, we clipped the log-likelihood matrix $\boldsymbol{L}$ of the existing 898 models at the lower 2nd percentile value -1495.9, following Oyama et al. (2025). The log-likelihood values of the newly added 120 models were also clipped using -1495.9 as the threshold.

To verify that this set of 10,000 texts is not biased, we sampled a new, entirely different set of 10,000 texts from the Pile and re-computed the distances (i.e., KL divergences) for all pairs among 100 models randomly sampled from the 1,018 models. The distances calculated on this new set showed a Pearson correlation of 0.997 with the distances from our original text set.

B.2 Visualization of Model Map

This section details the procedure for generating the t-SNE (van der Maaten and Hinton, 2008) visualizations and standard deviational ellipses presented in Fig. 1 and Fig. 3.

Coordinate Computation. First, the model coordinates $\widetilde{\boldsymbol{Q}}_d \in \mathbb{R}^{K\times d}$ are computed according to the method described in Section 3.2. These coordinates are subsequently reduced to two dimensions using the t-SNE algorithm. To assess variability of each resampling method, this entire procedure, from resampling to the final t-SNE mapping, is repeated $R = 100$ times. To ensure consistency across these trials, the stochastic t-SNE process is standardized: the initial coordinates for each run are determined by applying PCA to the model coordinate matrix $\widetilde{\boldsymbol{Q}}_d$ and the `random_state` parameter is fixed to 42.

Coordinate Alignment. Let $\boldsymbol{X}_r \in \mathbb{R}^{K\times 2}$ denote the matrix of t-SNE coordinates obtained from the r-th trial ($r = 1, \ldots, R$), and let $\boldsymbol{X}_{\text{ref}} \in \mathbb{R}^{K\times 2}$ be the reference coordinates derived from applying t-SNE to the matrix $\boldsymbol{Q} \in \mathbb{R}^{K\times N}$. To enable meaningful comparison across the R trials, we need to align these coordinate sets, as t-SNE results have inherent translational and rotational am-

[9] https://github.com/shimo-lab/modelmap/tree/main/1000models

	LS $d = 2{,}202$	Uniform $d = 6{,}320$	Uniform $d = 2{,}209$
Bias^2	3.19	3.00	4.95
Variance	4.65	4.23	10.83
MSE	8.22	7.42	17.48

Table 3: Bias^2, Variance, and MSE for model maps constructed using each sampling method. For each of 1,018 models, the values are calculated over $R = 100$ trials relative to the reference coordinate ($\boldsymbol{Y}_{\text{ref}}$) created from all $N = 10{,}000$ texts. The table reports the median values for these metrics across all 1,018 models.

biguities. The alignment process begins by centering each coordinate set: $\boldsymbol{Y}_r := \boldsymbol{X}_r - \bar{\boldsymbol{X}}_r$ and $\boldsymbol{Y}_{\text{ref}} := \boldsymbol{X}_{\text{ref}} - \bar{\boldsymbol{X}}_{\text{ref}}$. Subsequently, Orthogonal Procrustes analysis (Schönemann, 1966) is applied. For each $\boldsymbol{Y}_r$, we find an orthogonal transformation matrix $\boldsymbol{U}_r \in \mathbb{R}^{2\times 2}$ that best aligns it with $\boldsymbol{Y}_{\text{ref}}$. The final aligned coordinates are given by $\boldsymbol{Z}_r := \boldsymbol{Y}_r \boldsymbol{U}_r$.

Centrography. After aligning all $R = 100$ sets of coordinates, we analyze the variability of each model point. Let $\boldsymbol{z}_i^{(r)} \in \mathbb{R}^2$ be the 2D coordinate vector for model i from the r-th aligned set $\boldsymbol{Z}_r \in \mathbb{R}^{K\times 2}$. For each model coordinate, the mean coordinate $\bar{\boldsymbol{z}}_i = \frac{1}{R}\sum_{r=1}^{R} \boldsymbol{z}_i^{(r)}$ and the covariance matrix $\mathrm{Cov}(\boldsymbol{z}_i) \in \mathbb{R}^{2\times 2}$ are calculated across the R trials. The standard deviational ellipses (Yuill, 1971) shown in Fig. 1 and Fig. 3 are derived from these covariance matrices, with their height, width, and angle determined by the eigenvalues and eigenvectors of $\mathrm{Cov}(\boldsymbol{z}_i)$.

Quantitative Evaluation. We quantitatively evaluate the stability and accuracy of the model maps created using LS and Uniform sampling. The mean squared error (MSE) was calculated for each model i based on the aligned coordinate sets $\{\boldsymbol{Z}_r\}_{r=1}^{100}$ and the reference coordinate $\boldsymbol{Y}_{\text{ref}} = \left[\boldsymbol{y}_1^{\text{ref}}, \cdots, \boldsymbol{y}_K^{\text{ref}}\right]^\top \in \mathbb{R}^{K\times 2}$. The MSE is decomposed into bias, which indicates the accuracy, and variance, which reflects the stability of the resampling process:

$$\text{MSE} = \underbrace{\|\bar{\boldsymbol{z}}_i - \boldsymbol{y}_i^{\text{ref}}\|^2}_{\text{Bias}^2} + \underbrace{\frac{1}{R}\sum_{r=1}^{R} \|\boldsymbol{z}_i^{(r)} - \bar{\boldsymbol{z}}_i\|^2}_{\text{Variance}}$$

Table 3 shows the median values of these metrics across 1,018 models. The results indicate that LS sampling achieves comparable accuracy and stability than Uniform sampling with less than half the number of unique texts. Furthermore, when the number of unique texts is similar (approximately 2,200), LS sampling exhibits lower bias and variance, demonstrating that LS sampling can construct more accurate and stable maps.

C KL Sampling

Notation. We resample a text x_s from the dataset $D_N = \{x_1, \ldots, x_N\}$ with probability π_s. Let the n resampled texts be denoted by $\widetilde{D}_n = (x_{u_1}, \ldots, x_{u_n})$. Define $\widetilde{\boldsymbol{Q}} = (Q^{(u_1)}, \ldots, Q^{(u_n)}) \in \mathbb{R}^{K\times n}$, and $\widetilde{\boldsymbol{q}}_i \in \mathbb{R}^n$ as the i-th row vector of $\widetilde{\boldsymbol{Q}}$. Denoting the (i,t)-th element of $\widetilde{\boldsymbol{Q}}$ as $\widetilde{q}_i(x_{u_t})$, we see that this value is equal to $q_i(x_{u_t})$. Note that, unlike the notation used in Section 3.2, $\widetilde{\boldsymbol{q}}$ allows for duplication of the resampled texts and its columns are resampled from the double-centered matrix $\boldsymbol{Q}$.

Let $\boldsymbol{w}_n = (1/n\pi_{u_1}, \ldots, 1/n\pi_{u_n})^\top \in \mathbb{R}^n$ be the weights on the resampled texts, where

$$w_t = \frac{1}{n\pi_{u_t}}$$

for $t = 1, \ldots, n$, and let $\boldsymbol{W} = \mathrm{diag}(\boldsymbol{w}_n)$ be the corresponding diagonal matrix. Note that, unlike Section 3.2 where duplicate texts were summarized by counts $c(u_t)$, here duplicates are explicitly represented in $\widetilde{D}_n$, so $c(u_t)$ is not needed. Define $g_{ij} = \|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2$ and $\widetilde{g}_{ij} = \|\widetilde{\boldsymbol{q}}_i - \widetilde{\boldsymbol{q}}_j\|_{\boldsymbol{w}_n}$, where the weighted norm is taken with respect to $\boldsymbol{w}_n$.

Throughout Appendix C, the expectation operator $\mathbb{E}[\cdot]$ is taken with respect to the resampling procedure, conditional on the dataset D_N. That is, $\mathbb{E}[\cdot]$ should be interpreted as $\mathbb{E}[\cdot \mid D_N]$.

LS Sampling. According to Drineas and Kannan (2001), the expected Frobenius norm of the approximation error $\mathbb{E}\left[\|\widetilde{\boldsymbol{Q}}\boldsymbol{W}\widetilde{\boldsymbol{Q}}^\top - \boldsymbol{Q}\boldsymbol{Q}^\top\|_F^2\right]$ is minimized when the resampling probabilities satisfy $\pi_s \propto \|Q^{(s)}\|^2$.

KL sampling (Proposed). Instead of approximating the inner products in $\boldsymbol{Q}$, we aim to approximate the sum of the pairwise distances. We prove that the resampling probabilities π_s that minimize $\mathbb{E}\left[\sum_{i,j=1}^{K}(\widetilde{g}_{ij} - g_{ij})^2\right]$ are given by

$$\pi_s \propto \sqrt{\sum_{i,j=1}^{K} \left(q_i(x_s) - q_j(x_s)\right)^4}.$$

Lemma 1. For any $i, j \in \{1, \ldots, K\}$, it holds that

$$\mathbb{E}\left[\widetilde{g}_{ij}\right] = g_{ij}.$$

Proof. Since each resampled column is independently chosen from $\{1, \ldots, N\}$ with probability π_s, the expectation removes π_s and yields the full sum over all texts, as shown below:

$$\begin{aligned}
&\mathbb{E}[\widetilde{g}_{ij}] \\
&= \mathbb{E}[\|\widetilde{\boldsymbol{q}}_i - \widetilde{\boldsymbol{q}}_j\|_{\boldsymbol{w}_n}^2] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} \frac{1}{n\pi_{u_t}} (\widetilde{q}_i(x_{u_t}) - \widetilde{q}_j(x_{u_t}))^2\right] \\
&= \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\left[\frac{1}{\pi_{u_t}} (q_i(x_{u_t}) - q_j(x_{u_t}))^2\right] \\
&= \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{N} \frac{1}{\pi_s} (q_i(x_s) - q_j(x_s))^2 \pi_s \\
&= \|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2 \\
&= g_{ij}.
\end{aligned}$$

□

Lemma 2. The variance of the weighted distance after resampling is given by

$$\begin{aligned}
&\mathrm{Var}\,(\widetilde{g}_{ij}) \\
&= \frac{1}{n}\sum_{s=1}^{N} \frac{1}{\pi_s} \Big(q_i(x_s) - q_j(x_s)\Big)^4 - \frac{1}{n}\|\boldsymbol{q}_i - \boldsymbol{q}_j\|^4.
\end{aligned}$$

Proof. Since the n resampled columns are independent, the variance of the sum equals the sum of the variances of each term. Thus, by expanding the variance of each term and taking expectations with respect to π_s, we obtain

$$\begin{aligned}
&\mathrm{Var}(\widetilde{g}_{ij}) \\
&= \mathrm{Var}\left(\sum_{t=1}^{n} \frac{1}{n\pi_{u_t}} \Big(\widetilde{q}_i(x_{u_t}) - \widetilde{q}_j(x_{u_t})\Big)^2\right) \\
&= \sum_{t=1}^{n} \mathrm{Var}\left(\frac{1}{n\pi_{u_t}} \Big(\widetilde{q}_i(x_{u_t}) - \widetilde{q}_j(x_{u_t})\Big)^2\right) \\
&= \sum_{t=1}^{n} \left(\frac{1}{n^2}\sum_{s=1}^{N} \frac{1}{\pi_s} \Big(q_i(x_s) - q_j(x_s)\Big)^4 \right. \\
&\qquad \left. - \Big(\frac{1}{n}\|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2\Big)^2\right) \\
&= \frac{1}{n}\sum_{s=1}^{N} \frac{1}{\pi_s} \Big(q_i(x_s) - q_j(x_s)\Big)^4 - \frac{1}{n}\|\boldsymbol{q}_i - \boldsymbol{q}_j\|^4.
\end{aligned}$$

□

Proposition 1. The resampling probabilities π_s that minimize $\mathbb{E}\left[\sum_{i,j=1}^{K} (\widetilde{g}_{ij} - g_{ij})^2\right]$ are given by

$$\pi_s \propto \sqrt{\sum_{i,j=1}^{K} (q_i(x_s) - q_j(x_s))^4}.$$

Proof. Let $A_s = \sum_{i,j=1}^{K} \big(q_i(x_s) - q_j(x_s)\big)^4$ for $s = 1, \ldots, N$. By Lemma 1, we have

$$\mathbb{E}\left[\sum_{i,j=1}^{K} (\widetilde{g}_{ij} - g_{ij})^2\right] = \sum_{i,j=1}^{K} \mathrm{Var}(\widetilde{g}_{ij}).$$

Therefore, by Lemma 2, we aim to minimize

$$f(\pi_1, \ldots, \pi_N) = \sum_{s=1}^{N} \frac{1}{\pi_s} A_s$$

subject to the constraint $\sum_{s=1}^{N} \pi_s = 1$. Let

$$\begin{aligned}
&g(\pi_1, \ldots, \pi_N, \lambda) \\
&\quad = f(\pi_1, \ldots, \pi_N) + \lambda \left(\sum_{s=1}^{N} \pi_s - 1\right).
\end{aligned}$$

By setting $\dfrac{\partial g}{\partial \pi_s} = 0$, we obtain

$$\lambda \pi_s^2 = A_s.$$

Solving this with respect to π_s, we have

$$\pi_s = \frac{\sqrt{A_s}}{\sum_{s'=1}^{N} \sqrt{A_{s'}}}.$$

When $\pi_s > 0$,

$$\nabla^2 f(\pi_1, \ldots, \pi_N) = \mathrm{diag}\left(\frac{2A_1}{\pi_1^3}, \ldots, \frac{2A_N}{\pi_N^3}\right).$$

This Hessian is positive semidefinite, and positive definite if $A_s > 0$ for all s. The feasible set $\{\boldsymbol{\pi} : \sum_s \pi_s = 1,\ \pi_s \geq 0\}$, called the probability simplex, is convex. Since f is convex on this simplex (strictly convex when $A_s > 0$ for all s), the stationary point obtained above (with $\pi_s = 0$ whenever $A_s = 0$) is the global minimizer under the constraint.

□

D Estimating the Population Error of Resampled Distances

This section provides a detailed discussion of the error evaluation method introduced in Section 4.1.

Sampling Error. We consider the case where a dataset $D_n = (x_1, \ldots, x_n)$ of size n is sampled from a large population of texts $D^\dagger = (x_1^\dagger, \ldots, x_{N_0}^\dagger)$. Here we use n (instead of N used in the main text) to denote a general sample size, in order to analyze the error as a function of the number of sampled texts. We assume that $N_0 \gg n$ and n is sufficiently large. Let $\boldsymbol{L}^\dagger \in \mathbb{R}^{K \times N_0}$ and $\boldsymbol{L} \in \mathbb{R}^{K \times n}$ denote the log-likelihood matrices evaluated on $D^\dagger$ and D_n, respectively. Their doubly centered versions are denoted by $\boldsymbol{Q}^\dagger = [\boldsymbol{Q}^{\dagger(1)}, \ldots, \boldsymbol{Q}^{\dagger(N_0)}]$ and $\boldsymbol{Q} = [\boldsymbol{Q}^{(1)}, \ldots, \boldsymbol{Q}^{(n)}]$, with row vectors $\boldsymbol{q}_i^\dagger \in \mathbb{R}^{N_0}$ and $\boldsymbol{q}_i \in \mathbb{R}^n$ representing the coordinates of model p_i in each case. Since n is sufficiently large, the average vector used in centering $\boldsymbol{Q}$ can be well approximated by that of $\boldsymbol{Q}^\dagger$, and thus each column of $\boldsymbol{Q}$ may be regarded as a random sample drawn without replacement from the columns of $\boldsymbol{Q}^\dagger$.

We focus on the squared Euclidean distance between two models p_i and p_j. To ensure comparability with the case where the data size is $n = N$, we introduce a scaling factor to normalize the estimate and define:

$$\begin{aligned} g_{ij} &= \frac{N}{n} \|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2 \\ &= \frac{N}{n} \sum_{s=1}^{n} \left(q_i(x_s) - q_j(x_s)\right)^2, \\ g_{ij}^\dagger &= \frac{N}{N_0} \|\boldsymbol{q}_i^\dagger - \boldsymbol{q}_j^\dagger\|^2 \\ &= \frac{N}{N_0} \sum_{s=1}^{N_0} \left(q_i^\dagger(x_s^\dagger) - q_j^\dagger(x_s^\dagger)\right)^2. \end{aligned}$$

We define the sampling error as

$$\varepsilon_{ij} = g_{ij} - g_{ij}^\dagger.$$

Since the terms $\{(q_i(x_s) - q_j(x_s))^2\}_{s=1}^n$ can be regarded as random samples from $\{(q_i^\dagger(x_s^\dagger) - q_j^\dagger(x_s^\dagger))^2\}_{s=1}^{N_0}$, g_{ij} is an unbiased estimator of $g_{ij}^\dagger$, and thus the sampling error satisfies $\mathbb{E}[\varepsilon_{ij}] = 0$.

The mean squared error (MSE) of this sampling error is then given by

$$\kappa_{ij,n}^2 = \mathbb{E}[\varepsilon_{ij}^2].$$

Following the formulation in Section 4.1, we then aggregate these errors across all model pairs by considering their relative magnitudes:

$$\kappa_n^2 = \frac{1}{K^2} \sum_{i,j=1}^{K} \frac{\kappa_{ij,n}^2}{\max(g_{ij}^\dagger, \varepsilon_0)^2},$$

where $\varepsilon_0 > 0$ is a small constant to avoid division by zero.

Bootstrap Estimate of Sampling Error. We consider a bootstrap procedure that randomly samples n texts uniformly with replacement from the dataset D_n (we note that D_n may later be replaced by D_N). That is, each text in D_n is selected with equal probability $\pi_s = 1/n$. Let $\widetilde{D}_n = (\widetilde{x}_1, \ldots, \widetilde{x}_n) = (x_{u_1}, \ldots, x_{u_n})$ denote the resampled dataset. The log-likelihood matrix and its doubly centered version for $\widetilde{D}_n$ are denoted by $\widetilde{\boldsymbol{L}} \in \mathbb{R}^{K \times n}$ and $\widetilde{\boldsymbol{Q}} \in \mathbb{R}^{K \times n} = [\widetilde{\boldsymbol{Q}}^{(1)}, \ldots, \widetilde{\boldsymbol{Q}}^{(n)}]$, respectively, and the coordinate vector of model p_i is denoted by $\widetilde{\boldsymbol{q}}_i \in \mathbb{R}^n$. When n is sufficiently large, the mean vector used for centering $\widetilde{\boldsymbol{Q}}$ can be well approximated by that of $\boldsymbol{Q}$, so each column of $\widetilde{\boldsymbol{Q}}$ may be regarded as a random sample (with replacement) from the columns of $\boldsymbol{Q}$.

Unlike in Section 3.2, where resampling with replacement was handled by recording the number of duplicates and incorporating them as weights, we here represent the resampled data explicitly by allowing duplicate entries in the coordinate vectors $\widetilde{\boldsymbol{q}}_i$. Thus, the difference is only notational, and both approaches describe the same underlying resampling process.

As in the previous subsection, we focus on the squared Euclidean distance between models p_i and p_j. To ensure comparability with the case of dataset size $n = N$, we scale the estimate as follows:

$$\begin{aligned} \widetilde{g}_{ij} &= \frac{N}{n} \|\widetilde{\boldsymbol{q}}_i - \widetilde{\boldsymbol{q}}_j\|^2 \\ &= \frac{N}{n} \sum_{s=1}^{n} \left(\widetilde{q}_i(\widetilde{x}_s) - \widetilde{q}_j(\widetilde{x}_s)\right)^2. \end{aligned}$$

The terms $\{(\widetilde{q}_i(\widetilde{x}_s) - \widetilde{q}_j(\widetilde{x}_s))^2\}_{s=1}^n$ can be viewed as a bootstrap sample drawn (with replacement) from the set $\{(q_i(x_s) - q_j(x_s))^2\}_{s=1}^n$. We define the resampling error as

$$\widetilde{\varepsilon}_{ij} = \widetilde{g}_{ij} - g_{ij},$$

which satisfies $\mathbb{E}[\widetilde{\varepsilon}_{ij} \mid D_n] = 0$ by construction. The conditional mean squared error (MSE) of $\widetilde{g}_{ij}$ given D_n is defined as

$$\tau_{ij,n}^2 = \mathbb{E}\left[\widetilde{\varepsilon}_{ij}^2 \mid D_n\right].$$

In practice, we estimate this quantity by performing R independent bootstrap trials and computing

$$\tau_{ij,n}^2 = \frac{1}{R} \sum_{r=1}^{R} \left(\widetilde{\varepsilon}_{ij}^{(r)}\right)^2,$$

where $\widetilde{\varepsilon}_{ij}^{(r)}$ is the resampling error from the r-th trial.

This quantity $\tau_{ij,n}^2$ serves as the bootstrap estimate of $\kappa_{ij,n}^2$ (Efron and Tibshirani, 1994):

$$\hat{\kappa}_{ij,n}^2 = \tau_{ij,n}^2.$$

We may also replace D_n with D_N; in that case, this procedure corresponds to an n-out-of-N bootstrap (Bickel and Sakov, 2008; Shimodaira, 2014), and $\tau_{ij,n}^2$ remains a valid estimator of $\kappa_{ij,n}^2$. Intuitively, replacing D_n by a larger dataset D_N does not alter the distributional behavior of the resampled dataset $\widetilde{D}_n$, because its properties are determined solely by the resample size n rather than the population size. Moreover, the resampling MSE obeys a standard scaling law $\tau_{ij,n}^2 \propto n^{-1}$, and can be approximated by the theoretical relation

$$\tau_{ij,n}^2 = \frac{N}{n}\tau_{ij,N}^2.$$

Following Section 4.1, we define the aggregated resampling MSE as

$$\tau_{\text{unif},n}^2 = \frac{1}{K^2}\sum_{i,j=1}^{K}\frac{\tau_{ij,n}^2}{\max(g_{ij},\varepsilon_0)^2}.$$

Here $\tau_{ij,n}^2$ is based on uniform resampling from D_N with equal probability. This corresponds to the uniform resampling case discussed in Section 4.1. The above discussion on $\tau_{ij,n}^2$ also applies to the aggregated quantity $\tau_{\text{unif},n}^2$, since it is simply the average of the pairwise relative errors. Accordingly, the bootstrap estimate of the sampling MSE κ_n^2 is given by

$$\hat{\kappa}_n^2 = \tau_{\text{unif},n}^2.$$

The scaling law yields the approximation

$$\tau_{\text{unif},n}^2 = \frac{N}{n}\tau_{\text{unif},N}^2.$$

Decomposition of Population Error. In the previous subsection, we assumed uniform resampling, where each text is drawn with equal probability. Here, however, we consider the general case of weighted resampling. We consider a two-stage sampling procedure: first, we obtain a dataset D_N by randomly sampling N texts from the population $D^\dagger$; then, we perform weighted resampling with replacement from D_N to obtain a smaller dataset $\widetilde{D}_n$ of size n. As described in Section 3, we consider three types of resampling weights: LS sampling, KL sampling, and uniform sampling.

The estimated squared distance based on the resampled data is defined as

$$\widetilde{g}_{ij} = \|\widetilde{\boldsymbol{q}}_i - \widetilde{\boldsymbol{q}}_j\|_{\boldsymbol{w}_n}^2.$$

Here we adopt the notation for the weights without $c(u_t)$, because in $\widetilde{D}_n$ duplicates are explicitly represented rather than summarized by counts (see Appendix C). Thus each resampled entry carries the weight $w_t = 1/(n\pi_{u_t})$, which incorporates the scaling factor N/n so that the estimate matches the scale for sample size N. In particular, for uniform sampling ($\pi_{u_t} = 1/N$), this weight exactly reduces to N/n.

We now analyze the error of the resampling estimator $\widetilde{g}_{ij}$ relative to the true value $g_{ij}^\dagger$ in the population. Define the error as

$$\varepsilon_{ij}^\dagger = \widetilde{g}_{ij} - g_{ij}^\dagger.$$

This error can be decomposed as

$$\begin{aligned}\varepsilon_{ij}^\dagger &= (\widetilde{g}_{ij} - g_{ij}) + (g_{ij} - g_{ij}^\dagger)\\ &= \widetilde{\varepsilon}_{ij} + \varepsilon_{ij},\end{aligned}$$

where $\widetilde{\varepsilon}_{ij}$ and ε_{ij} denote the resampling error and the sampling error, respectively.

Taking the expectation of the squared error, we obtain the decomposition of the MSE:

$$\begin{aligned}\sigma_{ij,n}^2 &= \mathbb{E}[(\varepsilon_{ij}^\dagger)^2]\\ &= \mathbb{E}[(\widetilde{\varepsilon}_{ij} + \varepsilon_{ij})^2]\\ &= \mathbb{E}[\mathbb{E}[\widetilde{\varepsilon}_{ij}^2 \mid D_N]]\\ &\quad + 2\mathbb{E}[\mathbb{E}[\widetilde{\varepsilon}_{ij} \mid D_N]\varepsilon_{ij}] + \mathbb{E}[\varepsilon_{ij}^2]\\ &= \mathbb{E}[\tau_{ij,n}^2] + \kappa_{ij,N}^2,\end{aligned}$$

since the resampling error satisfies $\mathbb{E}[\widetilde{\varepsilon}_{ij} \mid D_N] = 0$. Here, we define the conditional MSE

$$\tau_{ij,n}^2 = \mathbb{E}[\widetilde{\varepsilon}_{ij}^2 \mid D_N],$$

which in practice is estimated from resampling errors $\widetilde{\varepsilon}_{ij}^{(r)}$ obtained via weighted resampling from D_N. The term $\kappa_{ij,N}^2 = \mathbb{E}[\varepsilon_{ij}^2]$ can be estimated by the bootstrap MSE under uniform resampling of size N, denoted by $\tau_{ij,\text{unif},N}^2$. Accordingly, we estimate $\sigma_{ij,n}^2$ by

$$\hat{\sigma}_{ij,n}^2 = \tau_{ij,n}^2 + \tau_{ij,\text{unif},N}^2.$$

As in Section 4.1, we define the aggregated population MSE from $\sigma_{ij,n}^2$ by

$$\sigma_n^2 = \frac{1}{K^2}\sum_{i,j=1}^{K}\frac{\sigma_{ij,n}^2}{\max(g_{ij}^\dagger,\varepsilon_0)^2}.$$

Let τ_n^2 denote the aggregated resampling MSE, defined from $\tau_{ij,n}^2$ as

$$\tau_n^2 = \frac{1}{K^2} \sum_{i,j=1}^{K} \frac{\tau_{ij,n}^2}{\max(g_{ij}, \varepsilon_0)^2},$$

which is in the same form as $\tau_{\text{unif},n}^2$ in the previous subsection but allows the resampling to use LS, KL, or uniform weights. Substituting the estimate $\hat{\sigma}_{ij,n}^2$ into $\sigma_{ij,n}^2$ in the expression of σ_n^2 yields the estimate:

$$\hat{\sigma}_n^2 = \tau_n^2 + \tau_{\text{unif},N}^2.$$

Note that while σ_n^2 is defined using the true distances $g_{ij}^\dagger$ in the denominator, its estimator τ_n^2 replaces them by the empirical distances g_{ij} computed from D_N.

E Model Performance Prediction

We computed the log-likelihoods of the unique texts contained in the resampled data and, following Oyama et al. (2025), used these values to predict model performance. This section describes the experiments in detail.

E.1 Model Performance

Following Oyama et al. (2025), we used the Open LLM Leaderboard v1 (Beeching et al., 2023) as the source of model performance scores[10]. The leaderboard provides scores for the following six benchmark tasks[11]: AI2 Reasoning Challenge (ARC) (Clark et al., 2018), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021), TruthfulQA (Lin et al., 2022), Winogrande (Sakaguchi et al., 2019), and GSM8K (Cobbe et al., 2021).

In addition to these benchmark scores, we followed Oyama et al. (2025) and also predicted (i) the average across the six tasks (hereafter referred to as 6-TaskMean) and (ii) the mean log-likelihood $\bar{\ell}_i$ of the log-likelihood vector $\boldsymbol{\ell}_i \in \mathbb{R}^N$.

E.2 Dataset Configuration

The number of resampled texts was set to

$$n \in \{10, 20, \ldots, 90, 100, 200, \ldots, 900, \\ 1{,}000, 2{,}000, \ldots, 9{,}000, 10{,}000\}.$$

As explained in Section 3, resampling n texts from D_N yields $\widetilde{D}_n$, and from this resampled set we extract a set of d unique texts, $D_d^* = \{x_{u_1}, \ldots, x_{u_d}\}$. Using these d texts, we computed the log-likelihood matrix $\boldsymbol{L}_d \in \mathbb{R}^{K \times d}$ and then formed the doubly-centered matrix $\widetilde{\boldsymbol{Q}}_d = [\widetilde{\boldsymbol{q}}_1, \ldots, \widetilde{\boldsymbol{q}}_K]^\top \in \mathbb{R}^{K \times d}$ with scaling weights $\boldsymbol{w}_d = \left(\frac{c(u_1)}{n\pi_{u_1}}, \ldots, \frac{c(u_d)}{n\pi_{u_d}}\right)^\top \in \mathbb{R}^d$.

For each benchmark task, the dataset is given as $\{(\widetilde{\boldsymbol{q}}_1, v_1), \ldots, (\widetilde{\boldsymbol{q}}_K, v_K)\}$, where $\widetilde{\boldsymbol{q}}_i = (\widetilde{q}_i(x_{u_1}), \ldots, \widetilde{q}_i(x_{u_d}))^\top \in \mathbb{R}^d$ is the i-th row of $\widetilde{\boldsymbol{Q}}_d$ corresponding to language model p_i, and $v_i \in [0, 100]$ is its benchmark score.

E.3 Regression Formulation

As in Oyama et al. (2025), we adopted ridge regression to predict each benchmark score. The matrix of explanatory variables is

$$\widetilde{\boldsymbol{Q}}_d \boldsymbol{W}_d^{1/2} \in \mathbb{R}^{K \times d}, \tag{1}$$

where the diagonal matrix $\boldsymbol{W}_d^{1/2} \in \mathbb{R}^{d \times d}$ has $\sqrt{\frac{c(u_t)}{n\pi_{u_t}}}$ on its t-th diagonal entry[12]. Let $\boldsymbol{v} = (v_1, \ldots, v_K)^\top \in \mathbb{R}^K$ denote the vector of target variables. The objective function, parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$, is defined as

$$\mathcal{L}(\boldsymbol{\theta}) = \|\boldsymbol{v} - \widetilde{\boldsymbol{Q}}_d \boldsymbol{W}_d^{1/2} \boldsymbol{\theta}\|^2 + \alpha \|\boldsymbol{\theta}\|^2,$$

where $\alpha \in \mathbb{R}_{>0}$ is a hyperparameter controlling the strength of regularization.

E.4 Training Setup

We partitioned the set of models into five folds according to their model types, as defined in Oyama et al. (2025). We then trained the parameters and predicted the benchmark scores. Training was performed with `RidgeCV` from scikit-learn (Varoquaux et al., 2015). To account for randomness, we repeated the data split with five different random seeds. As the evaluation metric, we computed Pearson's correlation coefficient (r) between the predicted and true benchmark scores for each split and averaged the results.

[10] https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard

[11] Benchmark scores are available for 996 of the 1,018 models released by Oyama et al. (2025). For convenience, we denote this subset size by K throughout this section.

[12] We adopt $\widetilde{\boldsymbol{Q}}_d \boldsymbol{W}_d^{1/2}$ as the matrix of explanatory variables rather than $\widetilde{\boldsymbol{Q}}_d$ itself. Let $\boldsymbol{W}_d = \mathrm{diag}(\boldsymbol{w}_d)$, whose t-th diagonal entry is $\frac{c(u_t)}{n\pi_{u_t}}$. Since $\boldsymbol{W}_d^{1/2} \boldsymbol{W}_d^{1/2} = \boldsymbol{W}_d$, we have $\widetilde{\boldsymbol{Q}}_d \boldsymbol{W}_d^{1/2} (\widetilde{\boldsymbol{Q}}_d \boldsymbol{W}_d^{1/2})^\top = \widetilde{\boldsymbol{Q}}_d \boldsymbol{W}_d \widetilde{\boldsymbol{Q}}_d^\top$. Then Lemma 1 of Drineas and Kannan (2001) gives $\mathbb{E}[\widetilde{\boldsymbol{Q}}_d \boldsymbol{W}_d \widetilde{\boldsymbol{Q}}_d^\top] = \boldsymbol{Q}\boldsymbol{Q}^\top$. Thus, pre-multiplying by $\boldsymbol{W}_d^{1/2}$ preserves this desirable expectation while appropriately re-scaling the features.

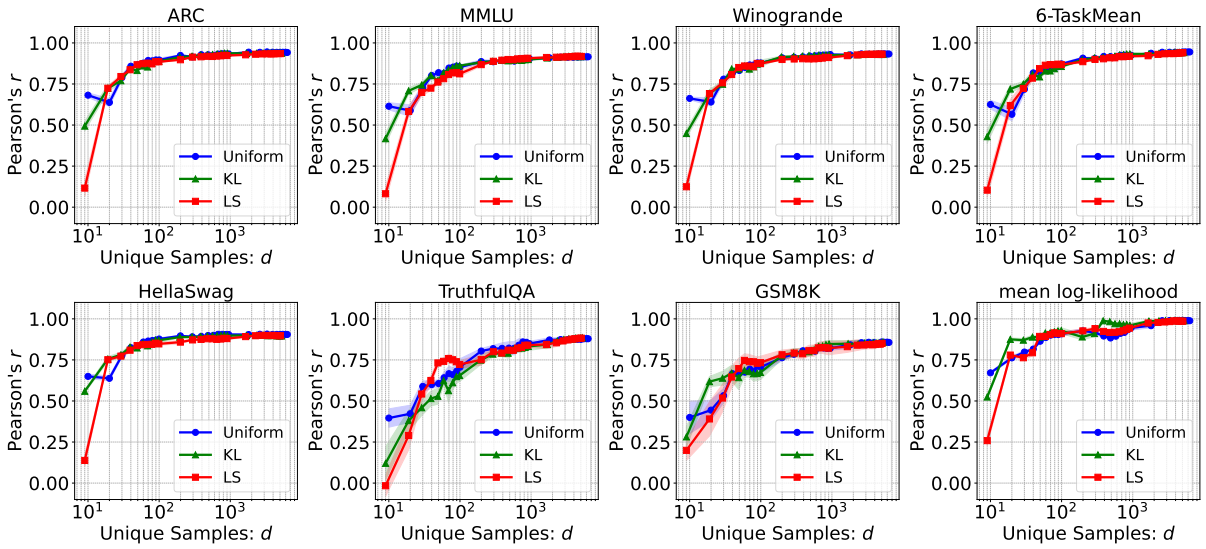


Figure 5: Pearson's correlation coefficient (r) between the predicted scores and the benchmark scores as a function of the number of unique texts d (determined by the resampling size n), plotted separately for each resampling method. Solid lines indicate the mean across five different data splits, and the shaded bands show ± 1 standard deviation. For every task and every method, predictive performance improves as d increases.

n	d	Method	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K	6-TaskMean	mean log-likelihood
10^1	10	Uniform	0.682 ± 0.014	0.650 ± 0.014	0.614 ± 0.022	0.396 ± 0.056	0.662 ± 0.008	0.400 ± 0.099	0.625 ± 0.026	0.671 ± 0.007
	9	KL	0.494 ± 0.029	0.559 ± 0.010	0.417 ± 0.034	0.120 ± 0.101	0.448 ± 0.025	0.281 ± 0.063	0.429 ± 0.041	0.524 ± 0.012
	9	LS	0.116 ± 0.051	0.138 ± 0.036	0.082 ± 0.045	-0.016 ± 0.065	0.124 ± 0.040	0.199 ± 0.061	0.104 ± 0.043	0.258 ± 0.030
10^2	100	Uniform	0.896 ± 0.012	0.877 ± 0.007	0.852 ± 0.012	0.705 ± 0.052	0.878 ± 0.007	0.706 ± 0.034	0.872 ± 0.006	0.909 ± 0.010
	99	KL	0.888 ± 0.012	0.870 ± 0.015	0.861 ± 0.006	0.655 ± 0.036	0.875 ± 0.021	0.674 ± 0.043	0.857 ± 0.008	0.928 ± 0.018
	99	LS	0.885 ± 0.005	0.847 ± 0.008	0.812 ± 0.016	0.722 ± 0.023	0.874 ± 0.009	0.733 ± 0.039	0.868 ± 0.006	0.910 ± 0.012
10^3	949	Uniform	0.933 ± 0.004	0.905 ± 0.005	0.903 ± 0.007	0.849 ± 0.015	0.929 ± 0.004	0.821 ± 0.012	0.925 ± 0.005	0.946 ± 0.014
	912	KL	0.937 ± 0.003	0.898 ± 0.006	0.899 ± 0.010	0.832 ± 0.025	0.926 ± 0.002	0.847 ± 0.022	0.933 ± 0.007	0.960 ± 0.012
	898	LS	0.923 ± 0.003	0.880 ± 0.004	0.906 ± 0.007	0.840 ± 0.012	0.911 ± 0.007	0.822 ± 0.041	0.919 ± 0.009	0.943 ± 0.015
10^4	6,335	Uniform	0.942 ± 0.002	0.905 ± 0.004	0.916 ± 0.006	0.879 ± 0.018	0.933 ± 0.005	0.857 ± 0.018	0.945 ± 0.004	0.989 ± 0.006
	5,240	KL	0.937 ± 0.002	0.896 ± 0.005	0.917 ± 0.005	0.885 ± 0.009	0.930 ± 0.003	0.857 ± 0.026	0.941 ± 0.005	0.988 ± 0.006
	5,080	LS	0.935 ± 0.002	0.897 ± 0.006	0.918 ± 0.007	0.882 ± 0.014	0.931 ± 0.005	0.851 ± 0.019	0.941 ± 0.004	0.986 ± 0.007

Table 4: Summary of the representative values from Fig. 5. For each resampling method, and for $n = 10^1, 10^2, 10^3, 10^4$ (with the corresponding numbers of unique texts d), the table reports Pearson's correlation r between the predicted and true benchmark scores, together with ± 1 standard deviation.

For each training set (i.e., the four folds in the outer five-fold CV), we conducted an inner five-fold cross-validation to select α from $\{10^1, \ldots, 10^9\}$, again following Oyama et al. (2025). The predicted scores were then clipped to the range $[0, 100]$. When the target variable $\boldsymbol{v}$ was the mean log-likelihood $(\bar{\ell}_1, \ldots, \bar{\ell}_K) \in \mathbb{R}^K$, we searched α in $\{10^{-4}, \ldots, 10^4\}$ and did not clip the predictions.

E.5 Results

Figure 5 shows Pearson's correlation coefficient between the predicted scores and the benchmark scores for Uniform, KL, and LS sampling as a function of the number of unique texts d for each resampling size n. Table 4 summarizes representative values obtained for each method at $n = 10^1, 10^2, 10^3, 10^4$ (and the corresponding d).

For all tasks and all methods, Pearson's correlation coefficient r increases as d grows. As shown in Table 4, even at $d \approx 100$, the predicted scores already achieve $r \approx 0.85$ under every resampling method, and only minor differences are observed among the strategies. Hence, predictive performance depends almost solely on the number of unique texts d.

This behavior can be interpreted as follows: when $d \ll K$ ($K \approx 10^3$), the column vectors of $\widetilde{\boldsymbol{Q}}_d \boldsymbol{W}_d^{1/2} \in \mathbb{R}^{K \times d}$ span a subspace of insufficient dimensionality, limiting the expressive power of the regression model. As d increases, the feature space expands and ridge regression becomes effective, leading to a rapid improvement in performance; however, once $d \gtrsim 10^3$ provides sufficient dimensionality, further gains in correlation are gradual.