Benchmark Distillation: Selecting Representative Evaluation Subsets via Component Relevance

Rajan Vivek
Stanford University

Kawin Ethayarajh
Stanford University
(Advisor)

Diyi Yang
Stanford University
(Advisor)

Douwe Kiela
Stanford University
(Advisor)

Abstract

While significant progress has been made in language model benchmarking, the sheer size and complexity of both the benchmarks and the models being evaluated makes it difficult for machine learning practitioners to grasp varying model behavior, edge cases, and weaknesses across the many dimensions encapsulated by any given benchmark. We propose Benchmark Distillation, a technique that leverages coreset strategies and the strong specialization of embedding components in modern language systems to identify small representative subsets of benchmarks. These subsets are designed to highlight model behavior across archetypal examples from the benchmark. We present a general recipe for achieving this vision and investigate a variety of strategies for each recipe component. We present preliminary results and propose a range of next steps.

1 Motivation

The incredible success of large pretrained models has marked a paradigm shift in machine learning from training single-task models to fine-tuning generalist models for specific downstream use cases. While a breadth of benchmarking strategies [14, 15, 16] have emerged to thoroughly and fairly assess model performance, their sheer size and complexity makes it difficult for machine learning practitioners to understand the strengths and weaknesses of various models across the dimensions encapsulated by the data. We aim to develop a technique for *benchmark distillation*, generating a small representative subset of a given benchmark such that 1) a given model achieves near equivalent performance on the subset and the entire benchmark as well as 2) the various behaviors of a given model on examples within the subset capture the breadth of behavior of the model on the entire benchmark. Such a technique would permit better diagnosis of model problems, identification of edge cases, and human awareness of the often non-intuitive behavior that these black-box systems exhibit. Our current work focuses on predicting downstream model behavior prior to fine-tuning, though future work may extend to assessing fine-tuned models as well.

2 Related Work

Our focus on evaluating *anticipated* downstream behavior of models prior to fine-tuning is similar in spirit to [1], which performs a large-scale investigation into predicting downstream task accuracy using upstream (pretraining) accuracy. Across thousands of experiments, they discover a nonlinear relationship reliably modeled with a power law curve, in which increasing upstream accuracy results in a saturation of downstream accuracy at a value dependent on the relatedness of the pretraining and downstream tasks. The curve appears for both fine-tuned downstream accuracy (where all layers are fine-tuned on 1000 downstream examples) and few-shot downstream accuracy (where all layers are frozen except a linear head trained on 1-25 downstream examples). We note that this work does not consider the optimal strategy for selecting representative downstream test examples.

The concept of selecting a small number of representative points from a dataset is well-explored in active learning and coreset literature. These strategies fall into two broad categories: uncertainty-based sampling and density-based sampling [3]. The former characterizes a dataset by selecting points close to decision boundaries, where a model tends to be uncertain. The latter characterizes a dataset by selecting points that evenly span the feature distribution such that all points are well-represented. However, [4] emphasize that selecting points for evaluation rather than learning warrants distinct selection criteria: 1) unlearnable points should still be selected, 2) selection bias should be removed (this bias tends to be helpful in active learning [5]), and 3) the variance of the evaluation loss estimate should be minimized. They propose a method of selecting high loss points using a surrogate model, leveraging the unbiased estimator for evaluation performance R_{PURE} from [5], which achieves minimal variance when the highest loss points are selected. We note that existing active selection literature does not consider the case where the dataset is composed of multiple disparate tasks.

The Logarithm of Maximum Evidence (LogME) metric was proposed by [9] to inexpensively predict downstream performance of a model during pretraining. The LogME metric assesses the compatibility between features F– data points encoded by the model– and corresponding labels by estimating the maximum value of label evidence given the features. This effectively measures the linear separability of F, which the authors show correlates strongly with fine-tuned performance. The authors do not propose a strategy for minimizing the amount of points necessary to achieve this strong correlation.

The findings of our embedding component analysis are consistent with [10], which found that zeroing high magnitude (outlier) activation components (accounting for 0.1% of the features overall) was detrimental to model performance, while zeroing a similar amount of randomly selected activation components had a negligible effect on model performance. This suggests that a small portion of embedding components are overwhelmingly responsible for model performance on a given task. We leverage this insight to select smaller subsets with minimal degradation in correlation with fine-tuned performance.

3 Methods and Experiments

Algorithm 1 Model Analysis via Representative Subset Evaluation

$D_{benchmark} \leftarrow \{\{$ example $x_{1,i}$, label $y_{1,i}\}_{i=1}^{L_1} ... \{$ example $x_{N,i}$, label $y_{N,i}\}_{i=1}^{L_N}\}$
$P \leftarrow$ point budget
$C \leftarrow$ component budget
$M \leftarrow$ candidate model
$D_{test} \leftarrow PointSelect(D_{benchmark}, P)$ ▷ select representative points
$F \leftarrow \mathrm{M}(D_{test}[x])$
$F' \leftarrow ComponentSelect(F, D_{test}[y], C)$ ▷ select representative components
$EvalScore \leftarrow GetLogME(F', D_{test}[y])$
$PointwiseScores \leftarrow PointWiseAnalysis(F', D_{test}[y])$ ▷ Get pointwise performance scores

3.1 Formulation

We aim to develop a procedure for evaluating and debugging model behavior using the general recipe shown above. In short, we propose a benchmark distillation strategy to compactly capture model performance and behavior across archetypal examples from the benchmark. The resulting representative evaluation set, which we refer to as a microset, requires selecting P examples from the benchmark (performed by $PointSelect$ in Algorithm 1) and C embedding components relevant to the task at hand (performed by $ComponentSelect$ in Algorithm 1). The microset is represented by a matrix of features F of size $P \times C$. Note that microsets are unique to each model and, depending on the benchmark distillation strategy, many microsets may exist for the same model. Microsets are designed such that the $LogME$ score of a model on the microset correlates strongly with the finetuned performance of the model on the benchmark.

3.1.1 Point Selection Strategies

Points should be selected that well-represent the overall distribution of characteristics exhibited by the data. We experiment with the following strategies:

1. **Random Selection**: Points are selected randomly, but the same points are selected for each model.
2. **K-Means Label-Aware Sampling**: A large number $N \gg P$ of examples from the benchmark are encoded by the model and K clusters are identified via K-Means. Next, $\frac{P}{K}$ points are sampled from each cluster.
3. **K-Means Label-Agnostic Sampling**: A large number $N \gg P$ of examples from the benchmark are encoded by the model and the encodings are segmented by label. For each of C labels, K clusters are identified via K-Means and $\frac{P}{K*C}$ points are sampled from each cluster.
4. **Herding Method [11]**: Points are selected based on the current subset mean and the original benchmark mean. Points are greedily added to the subset to minimize the difference between the two means.
5. **K-Center Greedy Method**: Points are selected to solve the minimax facility location problem [12]. Specifically, for all selected points the distance between it and its closest neighbor in the original dataset is minimized.

3.1.2 Component Selection Strategies

Embedding components that are most representative of information relevancy across all tasks should be chosen in order to overcome the embedding dimensionality bottleneck as described in Section 4.1. We experiment with the following strategies:

1. **Random Selection**: Components are selected randomly, but the same components are selected for each model.
2. **Principal Component Analysis (PCA)**: All embedding components are projected to the C component subspace that maximally preserves the variance in the original space.
3. **Maximum L1 Magnitude**: Following findings from [10] that suggest components with maximal magnitude are most relevant to model performance, we select the components with the largest L_1 magnitude.
4. L_1 **Magnitude Histogram Sampling**: We select points such that the distribution of L_1 magnitudes of the selected points matches the distribution of L_1 magnitudes across the original dataset.
5. **Maximum Variance**: We select components with the largest variance.
6. **Minimum Intra-Label Variance**: We select components with a metric rewarding minimal variance between points with the same label and large variance across points of all labels: $\frac{intra-label-variance}{overall-variance}$.

3.1.3 Pointwise Analysis Strategies

While the previous techniques work toward generating a microset whose LogMe score correlates strongly with overall model performance, gaining insight into model behavior on the benchmark requires observing model behavior at the individual example level. One promising approach towards this goal is to leverage the gradient of each point's label with respect to the microset LogME score. In theory, a point that is difficult for the model will exhibit a large positive gradient, suggesting that the model expects the point to be a different label. We intend to experiment with this in future work.

3.1.4 Evaluation

Combinations of point selection and component selection strategies define unique benchmark distillation techniques. To assess the quality of each technique, we pull 14 BERT-like models from HuggingFace Models [13] and generate microsets with each strategy for each model. For each of 5 benchmarks, we report the minimum microset size such that the Kendall Tau correlation between the LogMe scores of models on the distilled benchmark and LogME scores on the full benchmark surpasses 0.5, indicating that the rankings match with at least 0.75 probability. Note that we use the correlation with the overall LogME score (i.e. the LogME score computed on non-reduced encodings of $N \gg P$ points) as a proxy for correlation with downstream finetuned performance.

The models we used in our experiments are: "bert-base-uncased","roberta-base", "distilbert-base-uncased", "emilyalsentzer's Bio_ClinicalBERT", "dmis-lab's biobert-v1.1", "cardiffnlp's twitter-

roberta-base", "allenai's scibert_scivocab_uncased", 'gpt2', 'xlm-roberta-base', 'distilbert-base-uncased', 'albert-base-v2', 'funnel-transformer-small', 'distilbert-base-multilingual-cased', and 'distilroberta-base'.

The datasets we used in our experiments are SST-2 (movie sentiment classification), AG-News (article topic classification), Cola (linguistic acceptability classification), QNLI (question-answering natural language inference), and MNLI (multi-genre natural language inference).

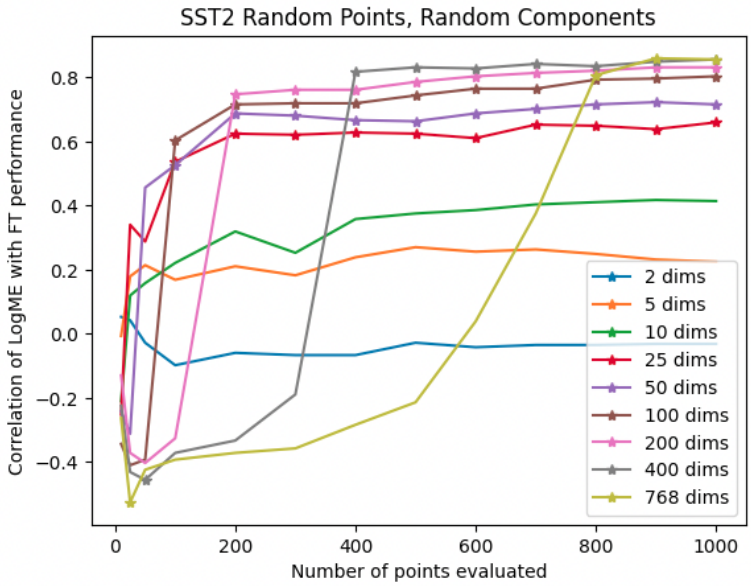

Figure 1: Correlation of various sized SST-2 microsets. We note that correlation drops as soon as the budget drops below the dimensionality. Points and components are selected randomly. Results are averaged over 5 runs.

4 Results and Discussion

4.1 The Embedding Dimensionality Bottleneck

Our first significant finding is that the microset size that sufficiently summarizes a model's encoding of a benchmark is lower-bounded by the dimensionality of that encoding. As shown in Figure 4.1, microset correlation with the overall LogMe score drops as soon $P < C$. This is likely related to the conditioning of $F^T F$, which is used to calculate LogME. We leave a formal proof explaining this finding for future work.

4.2 Representative Point Selection Strategies

We find that random point selection is a surprisingly strong baseline, achieving the smallest microset size for 3 of the 5 datasets evaluated. Herding selection achieves the smallest microset size for the remaining two datasets, but this performance is highly inconsistent: no microset of size less than 1000 is found via herding selection for Cola, despite random selection only requiring 200 points.

The weak performance of these techniques relative to random selection suggests a flaw in the general approach taken. One possible direction is to develop model-agnostic selection strategies, for example using a single large model to embed and sample points for all microsets. Another possible direction is to conduct component selection before rather than after point selection. It is likely that irrelevant components are influencing the point selection strategies.

4.3 Identifying Relevant Components

We present two significant findings regarding component statistics that relate to relevance. Firstly, we find that embedding component magnitude has a strong exponential relationship with component relevance, as measured by LogME. In Figure 3, we see that increasing component magnitude results in increasing LogME score. No such relationship was observed for component variance or intra-label variance, suggesting that components' variance is unrelated to relevance.

Dataset	Random, Random	K-Center Greedy, Random	Herding, Random	K-Means (Label Aware, K = 4), Random	K-Means (Label Agnostic, K = 4), Random
SST-2	**50** (0.52)	**200** (0.63)	**400** (0.69)	**200** (0.69)	**200** (0.59)
Ag-News	**50** (0.62)	**200** (0.51)	**25** (0.53)	**100** (0.61)	**50** (0.61)
Cola	**200** (0.55)	**> 1000**	**> 1000**	**500** (0.56)	**500** (0.51)
QNLI	**300** (0.57)	**900** (0.57)	**600** (0.50)	**400** (0.53)	**500** (0.63)
MNLI	**500** (0.6)	**> 1000**	**400** (0.59)	**500** (0.53)	**700** (0.54)

Figure 2: Size of smallest microset necessary to attain Kendall Tau of 0.5 with overall LogME score for all point selection strategies. Components are selected randomly for all results. The red values indicate the smallest microset found for each dataset. All results are averaged over 5 runs.

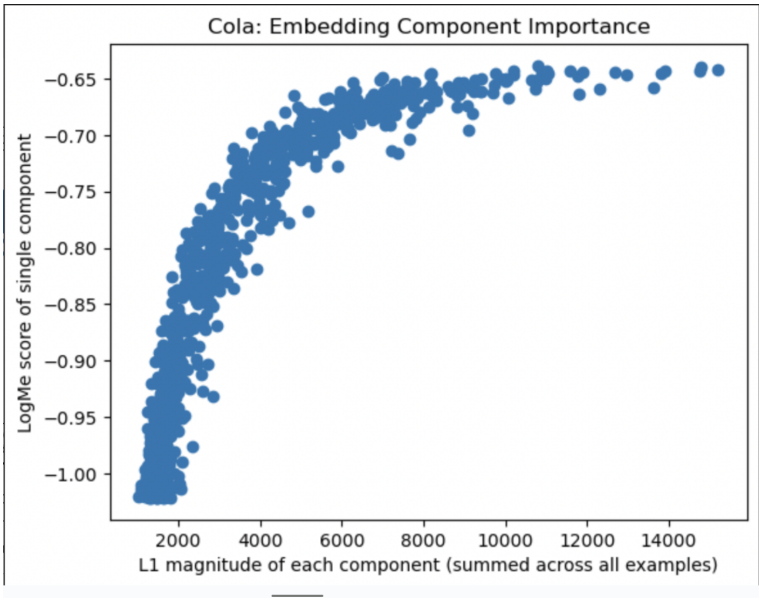

Figure 3: Exponential Relationship between Component Magnitude and Component LogME score for Cola. Near identical behavior was observed for other datasets.

Secondly, as shown in Figure 4, we find that pairs of components with small Pearson correlation magnitudes achieve greater LogME improvements over single components than more strongly correlated pairs. Specifically, a weak quadratic relationship is found where a Pearson correlation of zero maximizes average LogME improvement. This suggests that selecting informative components with non-redundant information better captures model performance.

4.4 Representative Component Selection Strategies

We observe that selecting components by minimum intra-label variance proves to be the strongest component selection strategy, finding the smallest microset for 3 of the 5 datasets and performing similarly to the best strategy on the remaining two datasets. This technique tends to outperform both PCA and maximum variance selection, suggesting that components with high variance are not worth selecting unless this variance is small across points with the same label.

We note that these results are counter-intuitive: selection by minimum intra-label variance strongly outperforms selection by maximum L1 magnitude despite only the latter exhibiting a strong relationship with component LogME score. These results suggest that simply selecting the most relevant components does not result in an optimal microset. We note that the improvement achieved by

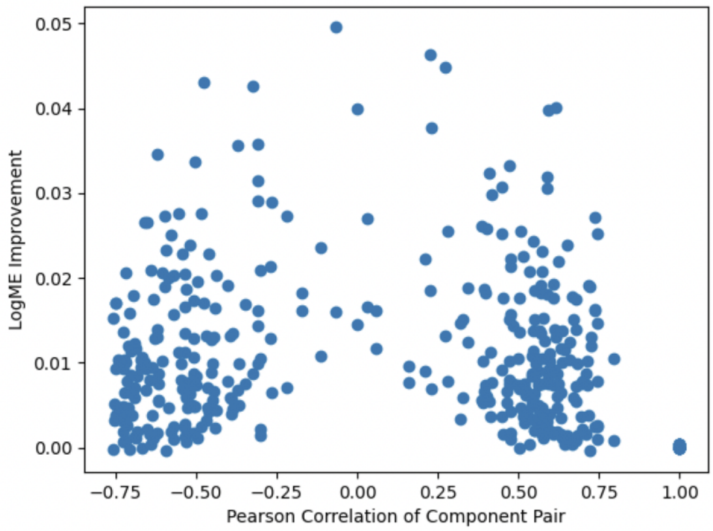

Figure 4: Weak Quadratic Relationship between Component Pair Pearson Correlations and LogMe improvement of joining the pair over considering one component. All pairs of components with the top 20 L1 magnitude are shown. The dataset is SST-2.

minimum intra-label variance selection may not be statistically significant, and intend to explore component selection techniques further.

Dataset	Random, Random	Random, PCA	Random, L1 Max	Random, Variance Max	Random, L1 Histogram	Random, Label Variance
SST-2	**50** (0.52)	**100** (0.65)	**100** (0.51)	**100** (0.51)	**100** (0.65)	**100** (0.74)
Ag-News	**50** (0.62)	**25** (0.55)	**100** (0.73)	**50** (0.65)	**50** (0.63)	**50** (0.63)
Cola	**200** (0.55)	**300** (0.65)	**300** (0.7)	**300** (0.65)	**200** (0.56)	**100** (0.52)
QNLI	**300** (0.57)	**800** (0.55)	**400** (0.57)	**400** (0.65)	**400** (0.63)	**200** (0.53)
MNLI	**500** (0.6)	**1000** (0.6)	**500** (0.57)	**500** (0.51)	**500** (0.53)	**500** (0.53)

Figure 5: Size of smallest microset necessary to attain Kendall Tau of 0.5 with overall LogME score for all component selection strategies. Points are selected randomly for all results. The red values indicate the smallest microset found for each dataset. All results are averaged over 5 runs.

5 Conclusion and Next Steps

We propose a recipe for benchmark distillation, where the performance and behavior of a model on a given benchmark can be compactly captured with a microset– a small representative evaluation subset. We find that the size of this microset is lower-bounded by the dimensionality of the embedding space, but selecting a relevant subset of embedding components can alleviate this constraint. The performance of various point and component selection strategies varies strongly for different datasets, suggesting that transformer models store various kinds of lingual information differently.

We hope to show that a single benchmark distillation technique can achieve strong performance across all datasets. Promising next directions are selecting points by only considering relevant embedding components rather than all components, identifying better component redundancy metrics, and modelling the distribution of useful components rather than simply selecting the most relevant C

components. This last direction is important because some models may have many more relevant components than others, resulting in a benchmark distillation technique with a fixed C underestimating the model performance.

Following the identification of a strong benchmark distillation technique, we hope to show that 1) label gradients or a similar method can approximate model behavior on individual examples and 2) the results of benchmark distillation can be visualized to improve user understanding of model behavior.

6 References

[1] Abnar, Samira, et al. "Exploring the limits of large scale pre-training." *arXiv preprint arXiv:2110.02095* (2021).

[2] Anonymous Authors. "The Role of Pretraining Data in Transfer Learning." *Under submission at ICLR 2023.*

[3] Emam, Zeyad Ali Sami, et al. "Active Learning at the ImageNet Scale." *arXiv preprint arXiv:2111.12880* (2021).

[4] Kossen, Jannik, et al. "Active testing: Sample-efficient model evaluation." International Conference on Machine Learning. PMLR, 2021.

[5] Farquhar, Sebastian, Yarin Gal, and Tom Rainforth. "On statistical bias in active learning: How and when to fix it." *arXiv preprint arXiv:2101.11665* (2021).

[6] Swayamdipta, Swabha, et al. "Dataset cartography: Mapping and diagnosing datasets with training dynamics." arXiv preprint arXiv:2009.10795 (2020).

[7] Achille, Alessandro, et al. "Task2vec: Task embedding for meta-learning." Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[8] Sener, Ozan, and Silvio Savarese. "Active learning for convolutional neural networks: A core-set approach." arXiv preprint arXiv:1708.00489 (2017).

[9] You, Kaichao and Liu, Yong. "LogME: Practical Assessment of Pre-trained Models for Transfer Learning" ICML 2021.

[10] Dettmers, Tim and Lewis, Mike and others. "LLM.init8(): 8-bit Matrix Multiplication for Transformers at Scale." NeurIPS 2022.

[11] Guo, Chengcheng and Zhao, Bo and Bai, Yanbing. DeepCore: A Comprehensive Library for Coreset Selection in Deep Learning." DEXA 2022.

[12] Farahani, Reza and Hekmatfar, Masoud. "Facility Location: Concepts, Models, Algorithms and Case Studies". 2009.

[13] HuggingFace Models. https://huggingface.co/models

[14] Wang, Alex and Singh, Amanpreet and others. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. ICLR 2019.

[15] Kiela, Douwe and Bartolo, Max and others. Dynabench: Rethinking Benchmarking in NLP. NAACL 2021.

[16] Liang, Percy and Bommasani, Rishi and others. Holisic Evaluation of Language Models. arxiv 2022.