Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Forecasting Fine-Tuning Break-Even Label Budgets in Text Classification from Frozen Embedding Geometry

Emin Talip Demirkiran

etd@eskisehir.edu.tr

Eskisehir Technical University

Research Article

Keywords: text classification, machine learning, label efficiency, zero-shot learning, sentence embeddings, class separability

Posted Date: September 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10975833/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Forecasting Fine-Tuning Break-Even Label Budgets in Text Classification from Frozen Embedding Geometry

Author: Emin Talip Demirkiran
Affiliation: Department of Computer Engineering, Eskisehir Technical University, Iki Eylul Campus, 26555 Eskisehir, Türkiye
Corresponding author: Emin Talip Demirkiran
Phone: +90 554 612 34 68
Email: etd@eskisehir.edu.tr
ORCID: https://orcid.org/0000-0001-6063-8178

Abstract

Choosing between zero-shot inference and supervised fine-tuning is often a label-budget decision: practitioners want to know when additional supervision will make adaptation worthwhile. Existing comparisons usually determine this threshold only after training models across multiple data budgets, leaving no low-cost way to anticipate it. We test whether class separability in frozen sentence-embedding spaces can serve as a training-free, although label-dependent, predictor of this break-even. Because separability itself requires class labels, the experiment deliberately evaluates a strong-form geometric signal rather than an annotation-free deployment rule. In a prospectively prespecified study spanning 35 public English-language text-classification tasks and three sentence encoders, we related pre-adaptation separability to the label budget at which a supervised Transformer classifier first exceeded a fixed instruction-tuned zero-shot comparator. Silhouette-based separability did not satisfy the prespecified inferential and practical-utility criteria, and an alternative separability measure produced a different directional pattern. Many tasks also failed to reach break-even within the tested budget, producing substantial right censoring. Censoring-aware, heteroskedasticity-robust, and influence-oriented post hoc analyses did not materially change the interpretation. Frozen embedding separability alone is therefore not a sufficiently stable estimator of fine-tuning break-even in this comparison regime. These findings place an empirical boundary on a plausible geometric heuristic and suggest that useful forecasts will likely require comparator-aware information, local boundary structure, or early adaptation dynamics.

Keywords: text classification; machine learning; label efficiency; zero-shot learning; sentence embeddings; class separability

1. Introduction

Text classification has repeatedly been reshaped by changes in representation learning and adaptation strategy. Classical supervised approaches were followed by neural architectures, pretrained Transformer encoders, parameter-efficient adaptation, prompt-based learning, and increasingly capable general-purpose language models [3, 8, 9, 11, 14, 21]. This expansion has created a practical model-selection problem that is not captured by conventional accuracy comparisons. For a new classification task, the relevant question is often not simply whether supervised fine-tuning can eventually outperform a prompted model, but whether the amount of labeled data needed to reach that point is small enough to justify annotation and training.

We call this quantity the fine-tuning break-even label budget: the smallest labeled training budget at which a task-specific classifier exceeds a fixed general-model comparator on the same held-out evaluation data. Recent comparisons show that this threshold can differ markedly between tasks and between comparison regimes. In particular, Pecher et al. [18] operationalized break-even performance between specialized small models and general large language models and showed that the required label budget is highly task dependent. Related studies of in-context learning, few-shot learning, and modern text-classification systems likewise show that the relative advantage of prompting and supervised adaptation depends on dataset, class structure, model family,

and available examples [4, 5, 20, 24, 25]. Yet these studies primarily *measure* the trade-off after experimentation. They do not provide a cheap diagnostic that forecasts it before a fine-tuning sweep is run.

A reliable forecast could reduce unnecessary fine-tuning sweeps and help structure annotation decisions, but the predictor studied here is not annotation-free. Class separability requires labeled examples; the present question is therefore whether a well-estimated, training-free geometric descriptor contains useful information about a supervised-versus-zero-shot crossing before the expensive fine-tuning sweep is run, not whether break-even can be predicted before any labels exist. We deliberately estimate separability from the full labeled task pool available for characterization, giving the geometric hypothesis a favorable measurement setting with less sampling noise. If even this strong-form signal is unstable, a deployable estimator based on a much smaller labeled pilot would require additional validation. From a knowledge-and-information-systems perspective, the broader problem remains one of data-acquisition decision support: determining which inexpensive task and system descriptors are informative enough to guide further annotation and adaptation.

The candidate descriptor is class separability in a frozen pretrained sentence-embedding space. Modern sentence encoders often reveal substantial semantic and task-relevant structure before downstream fine-tuning [6, 13, 17, 19, 23]. If examples from different classes are already geometrically well separated, fewer labeled examples might be needed for a supervised classifier to exploit that structure; conversely, strong class overlap might indicate a more difficult decision boundary and higher adaptation cost. The statistic is inexpensive with respect to model training but remains label dependent because class assignments are required to compute it. Across the literature reviewed for this study, we found no direct test of whether such pre-adaptation geometry forecasts a fine-tuning break-even label budget across a heterogeneous task collection.

The apparent simplicity of this hypothesis also makes it a useful target for rigorous falsification. A positive result would justify further development of low-cost geometric forecasting; a weak or unstable result would delimit where that shortcut is useful. We therefore designed the study around a prospectively prespecified confirmatory test rather than exploratory model selection. The comparison regime was fixed for reproducibility and mechanistic testing, not to benchmark contemporary model leadership. Before observing break-even outcomes, the hypothesis, primary separability measure, regression specification, directional test, practical-effect threshold, and decision rule were finalized and timestamped with cryptographic digests. The final design used 35 publicly available English-language text-classification tasks and three frozen sentence-embedding encoders.

The study asks one primary research question:

RQ1. Can class separability measured from labeled examples in a frozen sentence-embedding space, without task-specific training, predict the label budget at which supervised fine-tuning first outperforms a fixed zero-shot text-classification baseline?

The prespecified hypothesis was directional:

H1. Greater pre-adaptation class separability is associated with a lower fine-tuning break-even label budget after controlling for label count, dataset size, and mean input length.

The confirmatory analysis did not support H1 under the prespecified decision rule. Silhouette-based coefficients were negative for all three embedding encoders, as hypothesized, but none met both the inferential and practical-utility criteria. The largest observed partial association was still well below the prespecified large-effect threshold. More importantly, the result was not robust to the alternative geometric operationalization: a second, plausible inter-/intra-class distance ratio produced effects in the opposite direction. The experiment also revealed a structural limitation that is scientifically informative in its own right. Approximately two-thirds of the tasks were right-censored because the specialized classifier never exceeded the zero-shot baseline within the tested training grid.

These results motivate a more careful conclusion than either “embedding geometry predicts break-even” or “geometry is irrelevant.” The study establishes that the tested static separability signals are not sufficiently

strong or stable to support a stand-alone label-budget estimator in this regime. It also shows why break-even prediction should be treated as a censored, comparator-dependent resource-allocation problem rather than as an ordinary regression target. To strengthen that interpretation without altering the prespecified primary test, we add clearly labeled post hoc analyses for censoring, per-class scaling, heteroskedasticity, and leverage. None provides evidence that materially changes the primary interpretation.

The paper makes four contributions. First, it formalizes label-budget forecasting as prediction of a fine-tuning break-even point rather than as another end-state accuracy comparison. Second, it provides a prespecified multi-encoder test of frozen embedding-space separability as a training-free but label-dependent predictor across 35 heterogeneous tasks. Third, it documents how conclusions depend on the operational definition of separability and on heavy right censoring of the break-even outcome. Fourth, it converts a negative confirmatory result into constructive guidance for future forecasting work: models should explicitly accommodate censoring and are likely to require adaptation-dynamic, comparator-aware, or non-geometric task descriptors rather than relying on static class geometry alone.

2. Background and Related Work

Foundations and architecture evolution. The literature on text classification shows a progression from feature-based statistical learning to increasingly expressive learned representations. Broad surveys and system-level syntheses document this transition across classical machine learning, deep learning, pretrained language models, and graph-based methods [12, 21, 30, 31, 33, 71]. Early work established the importance of term weighting, Naive Bayes event models, support vector machines, and distributed word representations [34, 35, 37-39]. Neural approaches then shifted feature construction into the model itself through convolutional, hierarchical-attention, and graph-based architectures [10, 11, 40, 51, 63]. Contextualized representations and Transformer pretraining further changed the adaptation landscape through ELMo, the Transformer architecture, BERT, RoBERTa, ELECTRA, ALBERT, DistilBERT, XLNet, DeBERTa, and T5 [3, 15, 41-49]. Taken together, this line of work establishes that classification performance depends not only on dataset characteristics but also on the representation and adaptation mechanism used to exploit them.

Low-resource learning and efficient adaptation. A second line of research focuses on reducing the amount of task-specific supervision or adaptation required. Prompting surveys and foundational in-context-learning work describe how pretrained models can perform tasks from instructions or demonstrations without conventional fine-tuning [14, 44]. Domain- and task-adaptive pretraining, bottleneck adapters, and low-rank adaptation provide alternative routes for specializing pretrained models efficiently [7-9]. Few-shot methods such as pattern-exploiting training (PET), LM-BFF (Language Model-Better Few-Shot Fine-Tuning), and SetFit reduce labeled-data requirements through cloze reformulation, prompt/example construction, or contrastive sentence-embedding training [5, 20, 24], while parameter-free compression-based classification, dual-encoder intent models, synthetic-data-assisted parameter-efficient fine-tuning (PEFT), unseen-language adaptation, automated machine learning (AutoML)-guided fine-tuning, and intent-aware few-shot methods extend the low-resource design space [54, 59, 60, 61, 67, 75, 76]. Comparative studies of in-context learning and causal-LLM adaptation further show that the preferred strategy is conditional on the task, data regime, and adaptation formulation [4, 28]. This body of work motivates treating model selection as a resource-allocation problem rather than a single end-state accuracy comparison.

Sentence embeddings and representation quality. Sentence-embedding research provides the most direct conceptual basis for the predictor studied here. Sentence-BERT and SimCSE established efficient sentence-level representations and contrastive objectives for semantic similarity [6, 19], while MTEB provided a broad benchmark for evaluating text embeddings across heterogeneous tasks [17]. Instruction-conditioned and multi-stage contrastive approaches subsequently produced general-purpose encoders such as INSTRUCTOR and GTE [13, 23]. More recent work extends this trajectory through multilingual and multi-function embeddings, LLM-generated contrastive data, decoder-only LLMs converted into encoders, and repetition-based representation extraction [2, 22, 26, 65]. These results show that frozen representations can encode substantial semantic and task-relevant structure before downstream adaptation. They do not, however, imply that a global geometric statistic will predict the label budget needed for a different supervised model to overtake a particular

comparator. That distinction is central here: representation quality, label-cluster geometry, local decision-boundary structure, and adaptation speed are related but non-equivalent properties.

Large language model (LLM)-era classification and break-even trade-offs. The LLM-era classification literature makes the operational trade-off especially explicit. Synthetic-data studies examine when generated examples can supplement scarce or imbalanced labeled data [57, 62, 66], while entailment-based zero-shot classification and retrieval-augmented zero- and few-shot methods provide alternatives that avoid or postpone conventional fine-tuning [1, 27, 29]. Industrial and broad benchmark studies show that smaller specialized models can remain competitive and cost-effective despite the capabilities of general LLMs [64, 70]. Pecher et al. [18] most directly operationalize the break-even idea by asking how many labeled examples specialized models need to match or exceed general models; boundary-focused analysis [16] and multilingual LLM-era comparisons [25] further show that task structure and evaluation regime materially affect the answer. This work establishes break-even as a meaningful task-dependent quantity, but it generally estimates that quantity after running the competing systems across data budgets. It therefore leaves open whether a cheap pre-adaptation task descriptor can forecast the threshold in advance.

Robustness and non-standard task regimes. Finally, the broader evaluation literature cautions against assuming that findings from standard short-text, closed-set benchmarks generalize uniformly. Distribution-shift surveys and benchmarks, Robustness Gym, and empirical work on emotional-content shift demonstrate sensitivity to dataset and evaluation conditions [32, 50, 52, 56]. Long-document studies show that context length and architecture materially affect classification behavior [53, 55, 77], while preprocessing choices can still interact with model family [58]. Hierarchical and extreme multi-label classification, few-shot classifier chains, and open-world classification introduce substantially different label-space structures [36, 68, 69, 72, 73]; interpretability work also shows that LLM classification decisions may depend on latent concept structure not captured by simple global geometry [74]. Across the 77-source evidence base retained for this study, we found no prior work that directly tests whether frozen pre-adaptation class separability predicts a supervised-versus-zero-shot break-even label budget across a heterogeneous task collection. The present study addresses that narrower forecasting question while treating the answer as specific to the chosen comparator, supervised model, geometric measure, and tested label-budget range.

3. Method

3.1 Study design and prospective prespecification

The study was designed as a confirmatory test of one directional hypothesis. Before the break-even outcomes were inspected, the protocol specified the task collection, primary and secondary separability measures, models, training grid, break-even definitions, covariates, regression equation, one-sided significance level, practical-effect threshold, and decision rule. A time-stamped copy of the finalized protocol and resource specification was retained together with SHA-256 digests and an amendment log. Because the protocol was not deposited in an external public registry before execution, we describe the study as *prospectively prespecified* rather than *externally preregistered*.

A design-stage power calculation indicated that the initially contemplated task set would be inadequate for the prespecified large association. The task collection was therefore expanded to 35 before outcome analysis. For a one-sided test at $\alpha = 0.05$, the final sample size provided approximately 0.93 power to detect an association of magnitude $r = 0.50$. The design was thus intended to detect a large relationship with plausible operational value; it was not designed to exclude all small or moderate associations.

We distinguish three analysis classes. **Confirmatory** analyses were specified before outcomes were observed and determine the H1 decision. **Prespecified sensitivity analyses** were planned to probe robustness but do not alter that decision. **Post hoc** analyses were added after the censoring pattern became evident and are used only to aid interpretation.

3.2 Task corpus and data partitioning

The corpus contains 35 publicly available English-language text-classification tasks accessed through the Hugging Face `datasets` library [81]. Task selection was completed before any separability or break-even outcome was observed. An initial eight-task set was expanded after a pre-outcome power-adequacy review showed that it could not meaningfully support the prespecified four-predictor task-level regression. Candidate acquisition was purposive: it sought public single-label tasks that broadened variation in class count, acquired sample size, input length, domain, and classification type, while avoiding source-dataset overlap with the principal antecedent comparison studies [4, 18, 25]. Acquisition stopped at 35 after documented bounded technical attempts and diminishing returns. The corpus is therefore a deliberately heterogeneous benchmark sample, not a probability sample from a defined population of all text-classification tasks.

Task identities, class counts, source repositories, and original-source attributions where available are reported in Online Resource 1 (Table S1). The number of classes ranges from 2 to 151. After deterministic subsampling where necessary, available task sizes range from 635 to 3,000 examples, with a median of 3,000. Mean input length varies widely: the median is approximately 21 tokens, while one legal-document task has a mean length above 5,900 tokens and constitutes an extreme length outlier.

For each task, at most 3,000 examples were retained after a fixed-seed shuffle. A nominal 70% pool / 30% held-out evaluation split was then constructed with seed 42 while grouping exact duplicate texts so that identical content could not cross the boundary; consequently, realized split fractions could differ slightly from the nominal ratio. The evaluation split was never used to calculate separability or to construct fine-tuning subsamples. Separability was measured from the labeled pool, while the same held-out evaluation set was used to compare the supervised classifier and zero-shot baseline. This separation prevents direct reuse of evaluation labels in the predictor and outcome.

Task descriptors used as covariates were the number of classes, K; available task size after capping, N; and mean input length, L. These variables were selected before outcome analysis because they can influence both apparent geometry and the cost or difficulty of learning. In particular, class count is structurally important because the nominal total training budget scales with the number of classes.

3.3 Models and comparison regime

The specialized supervised model was bert-base-uncased with a freshly initialized task-specific classification head [3]. BERT was selected to keep the adaptation mechanism stable and reproducible rather than to maximize contemporary state-of-the-art performance. Fine-tuning used AdamW with learning rate 1e-5, 10% linear warm-up followed by linear decay, hidden dropout 0.3, 10 fixed epochs with no early stopping, and logical batch sizes between 4 and 32 according to the realized training-set size. Runs were repeated with seeds 42, 43, and 44; BERT text inputs were truncated to 256 tokens. The full optimizer, batching, sampling, model-version, and software-environment specification is given in Online Resource 1 (Table S2).

The general-model comparator was google/flan-t5-base, an instruction-finetuned T5 model [78], used in deterministic zero-shot mode. The exact prompt was `Classify the following text into one of these categories: {label_list}. Text: {text} Category:`. Generation used greedy decoding with no sampling or beam search; outputs were mapped to semantic label names by case-insensitive exact match followed by substring match, and unmatched outputs were scored as incorrect. Prompts were tokenized to a maximum total length of 512 tokens and generation was capped at 8 new tokens. No prompt optimization, demonstrations, retrieval augmentation, or model-specific task tuning was performed. We therefore treat FLAN-T5-base as a fixed prespecified zero-shot comparator, not as a claim about the strongest current general model. Exact model pins and parsing rules are reported in Online Resource 1 (Table S2).

Three frozen sentence encoders were used to calculate the candidate predictor: all-mpnet-base-v2, sup-simcse-roberta-base, and gte-base, representing sentence-transformer, SimCSE, and multi-stage contrastive embedding families [6, 13, 19]. all-mpnet-base-v2 and gte-base used their native mean-pooling configurations, whereas sup-simcse-roberta-base used the final-layer CLS representation. Encoder inputs were truncated to 256 tokens

and processed in batches of 32. All encoder parameters remained frozen; no outcome information entered the embedding computation.

One implementation deviation is relevant to interpretation. The protocol intended a uniform 256-token cap across models, but the executed FLAN-T5 code used a 512-token cap on the entire prompt, whereas BERT and the embedding encoders used 256-token text caps. Because the FLAN-T5 prompt also contains the instruction and full label list, these limits are not directly equivalent text budgets. The manuscript therefore characterizes the comparison actually executed and treats the unequal context cap as a threat to strict comparator symmetry rather than silently presenting the planned cap as the realized one.

3.4 Training-free class-separability measures

The primary separability statistic was the mean silhouette score [79], calculated with cosine distance using the task labels as cluster assignments. It was computed over the entire labeled pool P_t, not over a tiny pilot subset. The predictor is therefore training-free but label dependent; using the full pool deliberately gives the geometric hypothesis a comparatively favorable, low-measurement-noise test. The present experiment should not be interpreted as an annotation-free deployment procedure. For example i, the silhouette value contrasts its mean distance to members of the same class, a(i), with the smallest mean distance to any other class, b(i), through

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

The task-level statistic S(t,m) is the mean silhouette score across examples for task t under embedding model m. Larger values indicate tighter within-class cohesion relative to between-class distance. Negative values indicate that, on average, examples are closer to another class than to their assigned class under the embedding geometry.

A prespecified secondary separability measure used the ratio of average inter-class to intra-class cosine distance. The purpose of this measure was not to optimize prediction but to test whether any relationship was robust to a second reasonable operationalization of "separability." This distinction later proved important because the two measures produced qualitatively different regression behavior.

3.5 Fine-tuning break-even outcomes

The primary outcome B1(t) is the smallest nominal labeled training budget at which the mean macro-F1 of the fine-tuned BERT classifier exceeds the macro-F1 of the fixed FLAN-T5 zero-shot comparator for the same task.

The training grid used 1, 2, 4, and 8 labeled examples per class. For a task with K classes, these settings correspond to nominal total training budgets of K, 2K, 4K, and 8K examples. At each grid point and seed, examples were sampled separately without replacement within each class, taking `min(g, available examples in class)` at per-class grid value g; grid subsets were not constrained to be nested across budgets. The primary outcome is expressed in total labeled examples because total labels represent the annotation burden faced by a practitioner. Because this scale grows mechanically with K, class count is included as a primary covariate and B1/K is analyzed separately as a post hoc sensitivity.

If the fine-tuned classifier did not exceed the zero-shot comparator at any tested grid point, the task was recorded at the grid maximum and marked as right-censored. The prespecified primary OLS analysis uses this recorded value because that convention was fixed in advance. It should not, however, be interpreted as the true break-even point; the only valid statement is that the break-even exceeds the tested range.

A secondary outcome B2 uses a variance-aware criterion: the supervised model must exceed the baseline after subtracting one standard deviation from the supervised mean across the three repeats. B2 is therefore more conservative and is expected to be at least as large as B1.

3.6 Confirmatory regression and decision rule

For each embedding encoder m, we fit the prespecified ordinary least-squares (OLS) model

$$B_1(t) = \beta_0 + \beta_1 S(t,m) + \beta_2 K(t) + \beta_3 N(t) + \beta_4 L(t) + \varepsilon(t).$$

The four predictors S, K, N, and L were z-standardized before fitting; B1 was left on its original label-budget scale. Consequently, β1 should be interpreted as the expected change in B1 associated with a one-standard-deviation increase in the separability predictor, conditional on the other standardized covariates. It is a semi-standardized coefficient, not a fully standardized regression coefficient.

H1 predicts $\beta 1 < 0$. The prespecified inferential test was one-sided with $\alpha = 0.05$. Practical usefulness was evaluated using the partial correlation corresponding to the separability coefficient, with an absolute partial correlation of 0.50 or greater designated in advance as the minimum large effect that would justify treating the heuristic as operationally promising. The hypothesis was considered supported only if the directional inferential criterion and the practical-effect criterion were met consistently across the three encoders. Failure of either criterion was sufficient for rejection under the prespecified decision rule.

The one-sided test is appropriate to the locked directional hypothesis but does not prevent broader interpretation. We therefore report two-sided 95% confidence intervals for β1 and emphasize uncertainty rather than treating a near-threshold one-sided p-value as affirmative evidence. The confirmatory model retains the conventional OLS standard errors specified in advance; heteroskedasticity-robust standard errors and influence diagnostics are added post hoc and cannot change the confirmatory decision.

3.7 Prespecified robustness and sensitivity analyses

Several secondary analyses were planned to characterize the primary result without changing its decision status. The completed prespecified analyses reported here are:

1. **Alternative separability statistic (R1).** The inter-/intra-class distance ratio replaced silhouette score in the same regression specification.
2. **Variance-aware break-even (R2).** B2 replaced B1 as the outcome.
3. **Extreme input-length task (R3).** The legal-document task `lex_glue_scotus`, whose mean input length is far above the rest of the corpus and far beyond the 256-token BERT/embedding input cap, was excluded.
4. **Observed-break-even subset (R4).** The primary model was re-estimated only for tasks whose B1 was observed within the grid. Because this leaves only 13 tasks, it is descriptive and substantially underpowered.
5. **Label-permutation negative control (R6).** Labels were permuted before separability calculation to test whether an apparently predictive geometric signal survived destruction of label structure.

A planned exploratory few-shot prompting extension (R5) was not completed across its intended task subset because of resource constraints. No partial R5 result is analyzed or used in any conclusion; the incomplete extension is recorded solely as a protocol deviation.

The negative control produced an unexpected statistical association. Diagnostic analysis showed that the permuted silhouette statistic became strongly dependent on class count, especially for the three highest-K tasks. Because K also determines the nominal training-grid scale, this induced a structural artifact rather than providing evidence that the primary relationship was genuine. This diagnostic is reported because it exposes a useful caution: geometry statistics in many-class settings can acquire non-obvious dependencies on label cardinality even under random labels.

3.8 Post hoc censoring, scale, and influence analyses

The high censoring rate observed after execution raises a legitimate concern about treating the recorded grid maximum as if it were a fully observed continuous outcome. Because the primary analysis was prespecified, we do not replace it. Instead, four post hoc analyses are used to probe interpretation without altering H1's confirmatory status.

First, we fit a Cox proportional-hazards model [80] to B1, treating an observed break-even as the event and a grid-max non-break-even as right censored. The same standardized predictors S, K, N, and L are included. A hazard ratio greater than one for S is directionally consistent with earlier break-even at higher separability. With only 13 observed events, this model is used only as a sensitivity analysis.

Second, we analyze per-class break-even, B1/K, with the same covariate-adjusted OLS framework to check whether the primary result is dominated by the mechanical scaling of total label count with K.

Third, we inspect OLS residual and influence behavior and recompute the primary coefficient uncertainty using HC3 heteroskedasticity-robust standard errors. Breusch-Pagan tests, leverage, and Cook's distance are reported as diagnostics rather than as alternative confirmatory tests.

Fourth, because the three highest-cardinality tasks (K = 77, 100, and 151) are structurally unusual on the total-label outcome and were leverage-prone, we refit the primary regressions after excluding tasks with K >= 50. This threshold and analysis were not prespecified; they are explicitly diagnostic and are not used for hypothesis confirmation. Full results for the HC3 and high-K analyses are reported in Online Resource 1 (Table S3).

3.9 Reproducibility and computational environment

All 35 fine-tuning task evaluations and all 105 task-by-embedding separability computations completed successfully. The primary regression was independently recomputed from the stored task-level result table and reproduced the reported numerical output. Deterministic seeds were used throughout the supervised runs.

The experiments were executed with CPU-only training under a fixed software environment (Python 3.11.9; torch 2.13.0+cpu; transformers 5.14.1; sentence-transformers 5.6.1; scikit-learn 1.9.0; statsmodels 0.15.0; scipy 1.17.1). Fine-tuning and evaluation required approximately 3.25 hours of wall-clock time on the execution machine. Hardware affects runtime but is not treated as an experimental factor because model configurations and random seeds were held fixed. Online Resource 1 (Table S2) records the exact model identifiers, version pins, prompt/parsing rules, pooling choices, and training configuration.

Online Resource 2 supplies the prespecified protocol and prespecification record, environment information, standalone analysis reproduction code, derived task-level results, primary regression outputs, and the added diagnostic tables needed to inspect or reproduce the reported analyses. Third-party raw benchmark datasets and pretrained model weights are not redistributed; they should be reacquired from their original providers, with repositories and source attributions listed in Online Resource 1 (Table S1). Figure 1 was generated in Python using Matplotlib and is supplied separately as vector EPS artwork.

3.10 Use of generative AI tools

Generative AI tools were used under the author's supervision for literature organization, code drafting, statistical scripting, consistency checking, and language and formatting assistance. All scientific decisions, source verification, numerical validation, interpretation, and final approval were performed by the author. No generative AI system is listed as an author.

All numerical claims reported in the manuscript were checked against stored task-level outputs, and the primary regression was independently recomputed from those outputs. The author remains responsible for verifying the scientific reasoning and source evidence, approving the final manuscript, and accepting accountability for the integrity of the work.

4. Results

4.1 Descriptive characteristics and censoring

The task collection is deliberately heterogeneous. K ranges from 2 to 151 classes, while the median task has 3 classes. Most tasks contributed the capped maximum of 3,000 examples, although the smallest task contained 635. Input length is right-skewed: the median mean length is roughly 21 tokens, but one legal classification task has a mean above 5,900 tokens. This outlier is relevant because BERT and the embedding encoders use 256-token text caps; the executed FLAN-T5 comparator uses a 512-token cap on the full prompt, as disclosed in Sect. 3.3.

The zero-shot comparator also varies strongly across tasks, with macro-F1 ranging from near zero to above 0.94. This variation is important because B1 is a comparative outcome: the same supervised learning curve can produce a very different break-even depending on the strength of the zero-shot model on that task.

The most consequential descriptive feature is censoring. B1 is right-censored for 22 of 35 tasks (62.9%). In those tasks, the fine-tuned BERT classifier did not exceed the FLAN-T5 zero-shot comparator within the grid of up to 8 labeled examples per class. B2 is even more heavily censored: 28 of 35 tasks (80%) do not satisfy the variance-aware break-even criterion within the grid. Thus, the experiment provides exact break-even locations for a minority of tasks and lower bounds for the remainder.

The censoring pattern changes how the prespecified OLS coefficients should be interpreted. The recorded grid maxima are protocol-defined analysis values, not measured true break-even times. Claims about a universal continuous relationship beyond the tested budget would therefore be unwarranted.

4.2 Primary prespecified test

Table 1 reports the primary regression result for each frozen sentence encoder. All three β1 estimates have the hypothesized negative sign, meaning that greater silhouette-based separability is associated with a lower recorded break-even after adjustment for K, N, and L. However, none meets the prespecified criteria for support.

Table 1. Prespecified OLS association between silhouette separability and B1 (n = 35 for each encoder). Predictors are z-standardized; B1 remains in nominal labeled examples, so β1 is semi-standardized.

Encoder	β1 (SE)	95% CI	one-sided p	partial r	Decision
all-mpnet-base-v2	-41.19 (25.16)	[-92.58, 10.21]	0.056	-0.286	Not supported
sup-simcse-roberta-base	-19.53 (21.98)	[-64.41, 25.35]	0.191	-0.160	Not supported
gte-base	-17.32 (25.56)	[-69.52, 34.87]	0.252	-0.123	Not supported

Fig. 1 visualizes the estimated separability coefficients and their two-sided 95% confidence intervals.

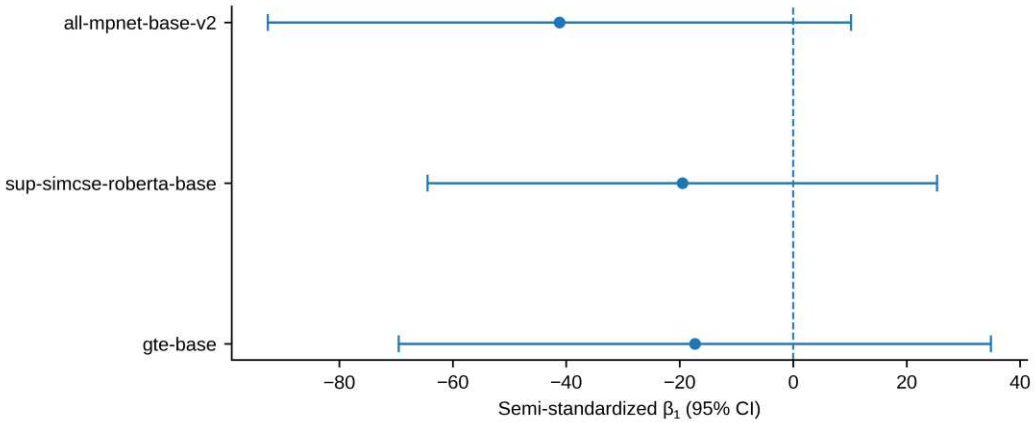


Fig. 1 Prespecified silhouette-separability coefficients for the three frozen sentence encoders; intervals are two-sided 95% confidence intervals

The all-mpnet encoder comes closest to the one-sided significance threshold, but the corresponding partial correlation is only -0.286, substantially below the prespecified practical threshold of $|r| = 0.50$, and its two-sided confidence interval includes both meaningful negative effects and small positive effects. The SimCSE and GTE estimates are weaker. Variance-inflation factors for primary separability are low (approximately 1.2-1.7), so the lack of support is not explained by severe multicollinearity in the primary models.

Under the prespecified decision rule, H1 is therefore not supported. More precisely, the data do not support the claim that frozen silhouette separability is a practically large, reliable stand-alone predictor of break-even across these tasks and encoders. The result should not be restated as proof of zero association: the study was powered for a large prespecified effect, and the confidence intervals still allow smaller effects.

4.3 Sensitivity to the definition of separability

The alternative distance-ratio statistic produced a qualitatively different pattern. Its $\beta 1$ estimate was positive for all three encoders, opposite to H1. The one-sided tests in the prespecified negative direction were correspondingly very far from significance. Partial correlations were positive, approximately 0.29-0.36. In addition, the alternative statistic was much more collinear with class count, with VIF values for the separability term around 9.7-17.6.

This sensitivity matters conceptually. If the result had been strong under both silhouette and distance ratio, one could argue for a relationship robust across reasonable geometric definitions. Instead, the sign changes when the operational definition of geometry changes. The most defensible conclusion is therefore not that "embedding geometry weakly predicts break-even," but that any apparent signal is metric-specific and unstable under another reasonable separability definition.

The divergence also suggests a methodological warning for future work. Aggregate class-geometry statistics can respond differently to label cardinality, class imbalance, within-class multimodality, and anisotropy of the embedding space. Treating "separability" as a single latent property without demonstrating measurement robustness may conceal these dependencies.

4.4 Variance-aware break-even

Replacing B1 with the variance-aware B2 leaves the broad conclusion unchanged. Coefficients remain negative for all three encoders, but the effect sizes and one-sided p-values do not meet the prespecified criteria. The all-mpnet encoder again has the largest-magnitude estimate ($\beta 1$ approximately -39.74, one-sided $p = 0.062$, partial $r = -0.277$), while the other two models are weaker.

The higher censoring rate of B2, 80%, is itself informative. Requiring the lower mean-minus-one-standard-deviation criterion of the supervised repeat distribution to exceed the zero-shot baseline substantially increases the label budget needed to claim a stable advantage. In practical deployment settings where variance matters, the inexpensive fine-tuning regime tested here often does not provide a clear win over the zero-shot comparator.

4.5 Extreme-length outlier

The task `lex_glue_scotus` has a mean input length of approximately 5,976 tokens, whereas the next-longest task is around 403 tokens. Because BERT and the embedding encoders truncate text at 256 tokens, this task is unusually mismatched to those components of the comparison regime; FLAN-T5 uses the separate 512-token full-prompt cap disclosed in Sect. 3.3. Removing it changes the all-mpnet result enough to cross the prespecified one-sided α threshold ($\beta 1 = -45.09$, $p = 0.047$), while SimCSE and GTE remain non-significant.

This exploratory exclusion does not alter the confirmatory conclusion. The exclusion was not part of the confirmatory decision; the all-mpnet partial correlation remains -0.306, well below the prespecified practical

threshold, and the pattern does not reproduce consistently across encoders. The analysis indicates that the extreme-length task exerts some leverage on the largest-magnitude primary estimate.

4.6 Observed-break-even subset

Excluding the 22 B1-censored tasks leaves only 13 tasks. In that subset, the separability coefficient becomes positive for all-mpnet and GTE and is near zero positive for SimCSE. None is significant in the hypothesized direction. Because this subset is small and selected on the outcome, it should not be treated as an unbiased estimator of the population relationship. It nevertheless reinforces the warning that the weak negative signal in the full prespecified OLS analysis is sensitive to how censored tasks are handled.

4.7 Label-permutation diagnostic

The label-permutation negative control initially yielded surprisingly strong negative coefficients rather than the expected null. Investigation showed that permuted silhouette values were highly correlated with K (Pearson correlations approximately -0.81 to -0.88 across encoders), with the strongest leverage concentrated in the three highest-class-count tasks. This is a structural property of the statistic under random multi-class partitions rather than evidence that meaningless labels genuinely forecast fine-tuning performance.

After excluding the three high-K leverage tasks, the apparent negative-control signal collapses, although the diagnostic regressions themselves become severely collinear. The episode is important because it reveals a failure mode that would be easy to overlook if only favorable controls were reported. It also motivates retaining K as a covariate and treating simple aggregate separability metrics cautiously in heterogeneous many-class corpora.

4.8 Post hoc censoring, scale, and influence analyses

Table 2 summarizes the censoring-aware and per-class analyses; the HC3 and high-cardinality diagnostics are reported in Online Resource 1 (Table S3). None of these post hoc analyses is part of the prespecified H1 decision.

Table 2. Post hoc analyses added to assess censoring and the K-scaled outcome. Cox HR > 1 is directionally consistent with earlier break-even at higher separability.

Encoder	Cox HR	two-sided p	B1/K β1	one-sided p	partial r
all-mpnet-base-v2	1.59	0.253	-0.464	0.259	-0.118
sup-simcse-roberta-base	1.00	>0.99	+0.295	0.686	+0.089
gte-base	1.19	0.626	-0.112	0.437	-0.029

In the Cox model, a hazard ratio above one is consistent with an earlier break-even at higher separability. All-mpnet and GTE point weakly in that direction, whereas SimCSE is essentially null, but all intervals are wide and all p-values are non-significant. Given only 13 observed events, the analysis is underpowered and should not be used to estimate precise survival effects. As a sensitivity analysis, explicitly modeling right censoring yields no material evidence of a stable association beyond the prespecified OLS result.

The per-class analysis likewise provides little evidence of a stable association. After expressing break-even as labeled examples per class, effects are small, inconsistent in sign, and non-significant. This reduces concern that the negative full-sample coefficient is merely a trivial consequence of total sample count scaling with K.

Additional OLS diagnostics indicate that conventional standard errors are fragile in this heterogeneous task sample. Breusch-Pagan tests reject homoskedastic residual variance for all three encoder regressions (all $p < 1e-5$), and the maximum leverage is approximately 0.994, concentrated on the extreme-length `lex_glue_scotus` task. Recomputing uncertainty with HC3 robust standard errors substantially widens the intervals: the one-sided p-values for all-mpnet, SimCSE, and GTE become 0.340, 0.348, and 0.396, respectively. The point estimates themselves remain the prespecified OLS estimates.

The high-cardinality diagnostic further weakens the apparent signal. After excluding the three tasks with K >= 50 (Banking77, LEDGAR, and CLINC150), the separability coefficients are -1.72, +0.37, and -1.10 for all-mpnet, SimCSE, and GTE, with corresponding partial correlations of -0.084, +0.019, and -0.055. Because this exclusion was selected post hoc for influence diagnosis, it is not evidence for a new effect; its value is to show that the weak full-sample negative slope is not stable once the most extreme label-cardinality tasks are removed. Online Resource 1 (Table S3) reports the complete HC3 and high-K diagnostic results.

Taken together, the confirmatory and post hoc evidence converges on the same operational conclusion: the tested frozen-embedding separability measures do not provide a sufficiently stable basis for stand-alone break-even forecasting.

5. Discussion

5.1 Interpreting the negative confirmatory result

The central finding is best understood as a boundary on a proposed decision heuristic, not as evidence that representation geometry is irrelevant to text classification. Frozen silhouette separability contains, at most, a weak and measurement-sensitive signal about the break-even outcome in this comparison regime. That is a materially narrower claim than saying that geometric structure has no predictive value. The study was designed to detect a large relationship that would be useful enough to support practical label-budget decisions; smaller associations remain compatible with the uncertainty in the estimates.

The deeper implication is that break-even is not an intrinsic property of a task. It is a relational outcome produced by the interaction of a task, a frozen task descriptor, a supervised learner, a zero-shot comparator, and a labeling schedule. A static representation can summarize how examples are arranged before adaptation, but it does not directly encode how quickly the supervised model will reorganize its representation or how high the comparator has already set the performance bar. This mismatch between a static predictor and a dynamic comparative target provides a more informative explanation for the weak result than simply labeling the predictor “bad.”

5.2 Why static geometry may be insufficient

Several mechanisms could weaken the connection between pre-adaptation geometry and break-even. Fine-tuning can substantially change task-relevant representations, so geometry measured before supervised adaptation need not resemble the geometry used by the final classifier [3, 7, 15]. A task that appears poorly separated in a frozen sentence space may still be learned quickly if a small amount of supervision induces useful representational change. Conversely, a task with clean frozen clusters may already be easy for the zero-shot comparator, delaying the point at which supervised adaptation becomes preferable.

The label dependence of the predictor further narrows its operational interpretation. We intentionally estimated separability from the entire labeled pool, which gives static geometry a favorable measurement setting and tests whether the signal exists before introducing the additional sampling noise of a small pilot. The present study therefore does not show that an annotation-free system can forecast label needs. A deployable version would need either a much smaller labeled characterization sample or a label-free proxy for geometry, and either choice would add a new source of uncertainty that should be validated prospectively.

Global cluster summaries may also miss the local structure that actually governs classification errors. Silhouette score averages cohesion and separation over the dataset, whereas a relatively small number of ambiguous boundary regions may dominate the learning problem. Boundary ambiguity has been identified as an important issue in LLM-era text classification [16], and robustness-oriented evaluation work likewise emphasizes the value of examining subpopulations and localized failure modes rather than relying only on aggregate statistics [50]. Future predictors may therefore benefit from neighborhood impurity, margin distributions, local density overlap, label-semantic similarity, or early representation drift rather than a single global separability score.

The disagreement between silhouette and the alternative distance ratio reinforces a measurement point: “class separability” is not a unitary construct. Compactness, centroid separation, nearest-class isolation, manifold overlap, and local margins can move differently across representations. A forecasting model should therefore begin with an explicit measurement model for geometry instead of treating several plausible statistics as interchangeable proxies for the same latent quantity.

5.3 Censoring changes the forecasting problem

The high proportion of tasks that did not reach break-even within the tested grid is not merely a nuisance in the regression; it reveals the natural statistical form of the target. Break-even behaves like a time-to-event variable in which the progression axis is labeled examples rather than clock time. When the experiment stops before the event occurs, the scientifically correct information is a lower bound, not an exact threshold. The Cox model used as a post hoc sensitivity analysis follows this logic [80], although the limited number of observed events prevents precise effect estimation.

This perspective suggests a better formulation for future systems. Instead of trying to predict one exact break-even count, a model could estimate the probability that supervised adaptation will overtake a comparator by a user-specified budget. Such a quantity is directly actionable—for example, whether break-even is likely by four labels per class—and naturally accommodates tasks whose threshold lies beyond the tested range. Prospective studies could use discrete-time hazards or accelerated-failure-time models and extend the training grid adaptively to obtain more informative event coverage.

5.4 Comparator dependence and model evolution

The chosen comparator matters as much as the task. Zero-shot and in-context performance vary with model family, prompt formulation, retrieval, and demonstrations [1, 4, 27, 29], while supervised adaptation behavior changes with architecture and fine-tuning strategy [8, 9, 24, 28]. Recent break-even and broad benchmark studies similarly show that relative performance shifts with both model scale and data regime [18, 25, 70]. The break-even observed here should therefore be read as the break-even under the specified BERT-versus-FLAN-T5 regime, not as a durable property of each dataset.

This dependence also explains why a task-only predictor is structurally incomplete. A useful forecasting system will likely need descriptors of both candidate systems: zero-shot confidence or calibration, prompt sensitivity, semantic alignment between labels and model outputs, early supervised learning-curve slope, and possibly representation change after a very small pilot update. As general-purpose and specialized models evolve, those relational features may transfer more naturally than a predictor that ignores the comparator.

5.5 Implications for practical label budgeting

For practitioners, the result argues against using frozen sentence-embedding silhouette score as a stand-alone rule for deciding whether to annotate and fine-tune. A more defensible low-cost strategy is staged. First, evaluate the chosen zero-shot comparator on a modest held-out labeled sample. Second, train the specialized model at the smallest one or two budgets and estimate the early learning-curve slope. Third, combine those dynamic signals with task descriptors such as class count, input length, imbalance, label-semantic similarity, and local embedding overlap. Few-shot methods and automated adaptation selection already demonstrate that useful decisions can be made from limited supervision without committing to a full training sweep [18, 24, 67].

The practical value of this staged approach is that it treats annotation as a sequential decision. Additional labels are acquired only when the expected information gain or performance gain justifies the cost. In that setting, frozen geometry may still contribute as one feature, but the present evidence does not justify letting it determine the decision alone.

5.6 What prespecified negative evidence contributes

The contribution of the negative result is methodological as well as empirical. A plausible heuristic was stated in advance, given a directional and practical-effect criterion, and tested across multiple embedding models. The resulting evidence does not support promoting that heuristic to a decision rule. This narrows the space of promising approaches and, importantly, exposes two design requirements that a purely positive benchmark might have obscured: the target must accommodate censoring, and the predictor must be robust to how geometric structure is operationalized.

The next step is therefore not to search for a slightly more favorable version of the same scalar relationship, but to test richer, explicitly comparator-aware forecasting models. Such models should separate confirmatory evaluation from feature exploration, define their censoring strategy prospectively, and assess whether any gain is large enough to change real annotation decisions.

6. Limitations and Threats to Validity

The study has several limitations that should be considered jointly rather than as isolated caveats.

Training-grid ceiling and heavy censoring. The largest tested budget is eight labeled examples per class. Because 22 of 35 B1 outcomes and 28 of 35 B2 outcomes are censored, the study cannot characterize the full break-even distribution beyond this regime. A larger grid may reveal later crossings and could change regression behavior. The post hoc survival analysis helps interpretation but is underpowered with only 13 observed B1 events.

Label dependence of the proposed predictor. Silhouette separability requires labeled class assignments and was estimated from the full labeled pool. The study therefore tests whether a well-estimated static geometric signal contains break-even information; it does not establish an annotation-free forecasting procedure. Estimating the same signal from a small pilot sample could introduce substantial measurement noise and requires separate validation.

Comparator dependence. The general baseline is FLAN-T5-base with a single fixed zero-shot prompt. Prompt optimization, stronger contemporary instruction models, retrieval augmentation, or few-shot demonstrations were not part of the confirmatory comparison. Break-even values should therefore be interpreted only for this comparator configuration; alternative prompting, retrieval, and instruction-model choices can materially change zero-shot performance [1, 4, 25, 29, 78]. Because the planned R5 few-shot extension was not completed and no partial R5 result is analyzed, the study does not support claims about prompted few-shot baselines.

Single supervised architecture. The specialized classifier is bert-base-uncased. The use of a stable, widely understood architecture improves internal comparability, but modern encoders and parameter-efficient adaptation methods may have different learning curves [8, 9, 24, 28]. A predictor could fail for BERT yet succeed for another adaptation family, or vice versa.

Embedding-model coverage. Three sentence encoders provide some replication across representations but do not span the rapidly expanding embedding landscape. Results may differ for larger LLM-based embeddings, task-adaptive embeddings, multilingual encoders, or models trained specifically for classification margins [2, 13, 17, 22, 23, 26, 65].

Operationalization of separability. Silhouette score is only one global measure of class geometry. The opposite-sign distance-ratio sensitivity demonstrates that conclusions are measurement dependent. Future work should compare local boundary measures, neighborhood impurity, class-centroid margins, density overlap, intrinsic dimensionality, and representation anisotropy under a prespecified measurement model.

Task-level sample size, purposive selection, and family dependence. A sample of 35 tasks provides strong power for the prespecified large-effect target but remains a modest sample for heterogeneous task-level

regression. The task set was purposively assembled to maximize variation and power rather than sampled probabilistically, and several tasks belong to related benchmark families (for example Yelp variants, GLUE-derived tasks, and LexGLUE subtasks). The 35 rows should therefore not be interpreted as 35 statistically independent draws from a defined population of all classification tasks. Covariate adjustment cannot account for all relevant task properties; domain, subjectivity, label semantics, class imbalance, annotation ambiguity, and distribution shift may remain unmeasured confounders.

Class-count structure and regression influence. Total nominal training size scales with K, and geometry statistics themselves can depend on K. K is explicitly controlled in the primary regression and the per-class sensitivity does not recover a signal, but the label-permutation diagnostic and high-K sensitivity show that extreme class cardinality can strongly affect task-level regressions. Post hoc residual diagnostics also show heteroskedasticity and very high leverage, so conventional OLS standard errors should not be read as the only uncertainty characterization; HC3-robust inference is reported as an additional sensitivity.

Context length and implementation deviation. BERT and the embedding encoders used 256-token text caps, which is severe for `lex_glue_scotus` and other long inputs [53, 55, 77]. The executed FLAN-T5 comparator instead used a 512-token cap on the complete prompt, despite the protocol's intended uniform 256-token cap. Because the prompt includes the instruction and label list, this does not translate to a simple twofold text-context advantage, but it does break strict context-length symmetry. The present results apply to the executed comparison and should not be interpreted as a perfectly context-matched benchmark.

Language and task scope. The corpus is English-language and dominated by single-sentence to short-document classification. Results should not be generalized to multilingual, extreme multilabel, hierarchical, open-set, or long-document classification without new evidence [25, 32, 36, 69, 72, 73, 77].

Benchmark and pretraining exposure. As with many studies using pretrained language models and public benchmarks, exact pretraining exposure to benchmark text or related corpora may be unknown. This can affect both the zero-shot comparator and frozen embedding geometry. The present design cannot fully audit training-corpus contamination across all pretrained models.

Macro-F1 and threshold definition. Break-even depends on the evaluation metric and on defining a crossing by mean macro-F1 over three repeats. Other metrics, confidence criteria, or repeated-measures designs could yield different thresholds. B2 partially addresses variance but becomes more heavily censored.

Post hoc sensitivity status. The Cox, per-class, HC3, and high-cardinality analyses were added after outcome inspection to address censoring, scale, and influence. Their convergence with the primary conclusion is reassuring but does not convert them into confirmatory evidence. A future study should prospectively specify censoring-aware modeling together with robust and influence-sensitive diagnostics from the outset.

7. Ethical and Reproducibility Considerations

The study uses pre-existing public benchmark datasets and does not recruit human participants or deploy a system to end users. No intervention, profiling decision, or automated high-stakes action was performed. The experiments nevertheless inherit any representational biases, annotation artifacts, or sensitive content present in the original benchmarks and pretrained models. These issues are particularly relevant for hate-speech, legal, medical, and social-media datasets and should be considered when interpreting task-level performance.

The principal societal implication is resource allocation rather than direct decision automation. A reliable break-even predictor could reduce unnecessary annotation and computation, whereas an unreliable predictor could cause teams to underinvest or overinvest in labeled data. The present result therefore supports caution against using a single embedding-space statistic as an unvalidated decision rule.

Reproducibility is supported by deterministic random seeds, a prospectively specified protocol, stored model identifiers, task-level derived results, complete primary and sensitivity outputs, and an independent recomputation of the primary regression. Online Resource 1 (Table S2) records the executed model, prompt,

pooling, training, and environment configuration, including the disclosed FLAN-T5 context-cap deviation; Online Resource 1 (Table S3) records the added HC3 and high-K diagnostics. Online Resource 2 contains the prespecified protocol and prespecification record, environment information, standalone analysis reproduction code, the 35-task derived result table, primary regression outputs, and the additional diagnostic tables. Raw third-party benchmark datasets and pretrained model weights are intentionally excluded and can be reacquired using the source information in Online Resource 1.

8. Conclusion

This study tested whether a well-estimated, training-free but label-dependent class-separability signal in frozen sentence-embedding spaces can forecast the label budget at which supervised fine-tuning overtakes a fixed zero-shot comparator. The prespecified evidence does not support frozen separability as a sufficiently strong and stable stand-alone estimator for that decision in the tested regime.

The central lesson is methodological rather than merely negative. The apparent association depends on how separability is operationalized, while the break-even outcome is heavily right-censored and intrinsically comparator dependent. These properties make break-even forecasting different from predicting end-state classification accuracy or generic task difficulty. A useful forecast must represent not only the task but also the competing systems and the dynamics of adaptation.

For practitioners, the present evidence favors small empirical pilots over static silhouette-based label-budget estimates. For researchers, it narrows the design space for future forecasting methods: censoring-aware models, comparator-strength descriptors, local boundary measures, and early learning-dynamic signals are stronger candidates than a single global frozen-geometry statistic. The contribution is therefore a clearer formulation of the forecasting problem and an empirical boundary on one plausible, training-free geometric heuristic rather than a claim of an annotation-free decision rule.

Statements and Declarations

Funding: No funding was received for conducting this study.

Competing interests: The author has no relevant financial or non-financial interests to disclose.

Author contributions: Emin Talip Demirkiran is the sole author of this manuscript and is responsible for the scientific conception and design, study implementation and oversight, interpretation and verification of the results, critical revision of the manuscript, and approval of the final version.

Ethics approval: Not applicable. This study consisted of secondary computational analysis of pre-existing publicly accessible benchmark datasets and involved no recruitment, intervention, interaction with human participants, collection of new human-participant data, or animal research.

Consent to participate: Not applicable.

Consent for publication: Not applicable.

Data availability: The source benchmark datasets are publicly available from their original providers and were accessed through the Hugging Face `datasets` library [81]. Dataset repositories and source attributions are listed in Online Resource 1 (Table S1). Third-party raw datasets and pretrained model weights are not redistributed. Derived task-level results, primary analysis outputs, the prespecified protocol and prespecification record, environment information, and supporting analysis materials are supplied in Online Resource 2.

Code availability: Standalone code for reproducing the task-level primary regressions and reported diagnostics from the included derived data is supplied in Online Resource 2.

Use of generative AI: Generative AI assistance used during the research and manuscript-preparation process is described in Sect. 3.10. The author is responsible for the scientific content, source verification, interpretation, final review, and approval of the manuscript.

Supplementary Information: Online Resource 1 (`ESM_1.pdf`) contains Tables S1-S3: task corpus and source attribution, the executed model/prompt/training configuration, and post hoc regression diagnostics. Online Resource 2 (`ESM_2.zip`) is a reproducibility package containing the prespecified protocol and prespecification record, environment information, standalone analysis reproduction code, derived task-level results, and primary-analysis outputs; it contains no third-party raw datasets or pretrained model weights.

References

1. Abdullahi T, Singh R, Eickhoff C (2024) Retrieval augmented zero-shot text classification. In: Proceedings of the 2024 ACM SIGIR International Conference on the Theory of Information Retrieval, pp 195-203. Association for Computing Machinery. https://doi.org/10.1145/3664190.3672514
2. BehnamGhader P, Adlakha V, Mosbach M, Bahdanau D, Chapados N, Reddy S (2024) LLM2Vec: large language models are secretly powerful text encoders. In: Conference on Language Modeling (COLM 2024). https://openreview.net/forum?id=IW1PR7vEBf
3. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional Transformers for language understanding. In: Proceedings of NAACL-HLT 2019, pp 4171-4186. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423
4. Edwards A, Camacho-Collados J (2024) Language models for text classification: is in-context learning enough? In: Proceedings of LREC-COLING 2024, pp 10058-10072. ELRA and ICCL. https://aclanthology.org/2024.lrec-main.879/
5. Gao T, Fisch A, Chen D (2021) Making pre-trained language models better few-shot learners. In: Proceedings of ACL-IJCNLP 2021, pp 3816-3830. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.295
6. Gao T, Yao X, Chen D (2021) SimCSE: simple contrastive learning of sentence embeddings. In: Proceedings of EMNLP 2021, pp 6894-6910. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.552
7. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, Smith NA (2020) Don't stop pretraining: adapt language models to domains and tasks. In: Proceedings of ACL 2020, pp 8342-8360. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.740
8. Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, de Laroussilhe Q, Gesmundo A, Attariyan M, Gelly S (2019) Parameter-efficient transfer learning for NLP. In: Proceedings of the 36th International Conference on Machine Learning. Proc Mach Learn Res 97:2790-2799. https://proceedings.mlr.press/v97/houlsby19a.html
9. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W (2022) LoRA: low-rank adaptation of large language models. In: International Conference on Learning Representations. https://openreview.net/forum?id=nZeVKeeFYf9
10. Joulin A, Grave E, Bojanowski P, Mikolov T (2017) Bag of tricks for efficient text classification. In: Proceedings of EACL 2017, pp 427-431. Association for Computational Linguistics. https://aclanthology.org/E17-2068/
11. Kim Y (2014) Convolutional neural networks for sentence classification. In: Proceedings of EMNLP 2014, pp 1746-1751. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1181
12. Kowsari K, Jafari Meimandi K, Heidarysafa M, Mendu S, Barnes L, Brown D (2019) Text classification algorithms: a survey. Information 10(4):150. https://doi.org/10.3390/info10040150
13. Li Z, Zhang X, Zhang Y, Long D, Xie P, Zhang M (2023) Towards general text embeddings with multi-stage contrastive learning. arXiv:2308.03281. https://doi.org/10.48550/arXiv.2308.03281
14. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G (2023) Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv 55(9):195. https://doi.org/10.1145/3560815
15. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692
16. Lu Z, Tian J, Wei W, Qu X, Cheng Y, Xie W, Chen D (2024) Mitigating boundary ambiguity and inherent bias for text classification in the era of large language models. In: Findings of ACL 2024, pp 7841-7864. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.467
17. Muennighoff N, Tazi N, Magne L, Reimers N (2023) MTEB: massive text embedding benchmark. In: Proceedings of EACL 2023, pp 2014-2037. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-main.148
18. Pecher B, Srba I, Bielikova M (2025) Comparing specialised small and general large language models on text classification: 100 labelled samples to achieve break-even performance. In: Proceedings of EMNLP 2025, pp 165-184. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.9

19. Reimers N, Gurevych I (2019) Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Proceedings of EMNLP-IJCNLP 2019, pp 3982-3992. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410
20. Schick T, Schütze H (2021) Exploiting cloze-questions for few-shot text classification and natural language inference. In: Proceedings of EACL 2021, pp 255-269. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.20
21. Sebastiani F (2002) Machine learning in automated text categorization. ACM Comput Surv 34(1):1-47. https://doi.org/10.1145/505282.505283
22. Springer JM, Kotha S, Fried D, Neubig G, Raghunathan A (2025) Repetition improves language model embeddings. In: International Conference on Learning Representations 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/e9b0c9480b03f9a6bca24ed8c08ee058-Abstract-Conference.html
23. Su H, Shi W, Kasai J, Wang Y, Hu Y, Ostendorf M, Yih WT, Smith NA, Zettlemoyer L, Yu T (2023) One embedder, any task: instruction-finetuned text embeddings. In: Findings of ACL 2023, pp 1102-1121. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.71
24. Tunstall L, Reimers N, Seo Jo UE, Bates L, Korat D, Wasserblat M, Pereg O (2022) Efficient few-shot learning without prompts. arXiv:2209.11055. https://doi.org/10.48550/arXiv.2209.11055
25. Vajjala S, Shimangaud S (2025) Text classification in the LLM era—where do we stand? arXiv:2502.11830. https://doi.org/10.48550/arXiv.2502.11830
26. Wang L, Yang N, Huang X, Yang L, Majumder R, Wei F (2024) Improving text embeddings with large language models. In: Proceedings of ACL 2024, pp 11897-11916. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.642
27. Yin W, Hay J, Roth D (2019) Benchmarking zero-shot text classification: datasets, evaluation and entailment approach. In: Proceedings of EMNLP-IJCNLP 2019, pp 3914-3923. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1404
28. Yousefiramandi A, Cooney C (2025) Fine-tuning causal LLMs for text classification: embedding-based vs. instruction-based approaches. arXiv:2512.12677. https://doi.org/10.48550/arXiv.2512.12677
29. Yu G, Liu L, Jiang H, Shi S, Ao X (2023) Retrieval-augmented few-shot text classification. In: Findings of EMNLP 2023, pp 6721-6735. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.447
30. Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J (2021) Deep learning-based text classification: a comprehensive review. ACM Comput Surv 54(3):62. https://doi.org/10.1145/3439726
31. Qiu X, Sun T, Xu Y, Shao Y, Dai N, Huang X (2020) Pre-trained models for natural language processing: a survey. Sci China Technol Sci 63(10):1872-1897. https://doi.org/10.1007/s11431-020-1647-3
32. Costache AV, Gheorghe SF, Poesina E, Irofti P, Ionescu RT (2026) Text classification under class distribution shift: a survey. In: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pp 4045-4060. Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.eacl-long.189
33. Wang K, Ding Y, Han SC (2024) Graph neural networks for text classification: a survey. Artif Intell Rev 57:190. https://doi.org/10.1007/s10462-024-10808-0
34. Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Inf Process Manag 24(5):513-523. https://doi.org/10.1016/0306-4573(88)90021-0
35. McCallum A, Nigam K (1998) A comparison of event models for Naive Bayes text classification. In: AAAI-98 Workshop on Learning for Text Categorization, pp 41-48
36. Liu R, Liang W, Luo W, Song Y, Zhang H, Xu R, Li Y, Liu M (2023) Recent Advances in Hierarchical Multi-label Text Classification: A Survey. arXiv:2307.16265. https://doi.org/10.48550/arXiv.2307.16265
37. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781. https://doi.org/10.48550/arXiv.1301.3781
38. Joachims T (1998) Text categorization with support vector machines: learning with many relevant features. In: Machine Learning: ECML-98. Lecture Notes in Computer Science, vol 1398, pp 137-142. Springer. https://doi.org/10.1007/BFb0026683
39. Pennington J, Socher R, Manning CD (2014) GloVe: global vectors for word representation. In: Proceedings of EMNLP 2014, pp 1532-1543. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162
40. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical attention networks for document classification. In: Proceedings of NAACL-HLT 2016, pp 1480-1489. Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1174

41. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: Proceedings of NAACL-HLT 2018, pp 2227-2237. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1202
42. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Advances in Neural Information Processing Systems 30, pp 5998-6008
43. Clark K, Luong MT, Le QV, Manning CD (2020) ELECTRA: pre-training text encoders as discriminators rather than generators. In: International Conference on Learning Representations. https://arxiv.org/abs/2003.10555
44. Brown TB, Mann B, Ryder N, et al (2020) Language models are few-shot learners. In: Advances in Neural Information Processing Systems 33, pp 1877-1901
45. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2020) ALBERT: a lite BERT for self-supervised learning of language representations. In: International Conference on Learning Representations. https://arxiv.org/abs/1909.11942
46. Sanh V, Debut L, Chaumond J, Wolf T (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108. https://doi.org/10.48550/arXiv.1910.01108
47. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV (2019) XLNet: generalized autoregressive pretraining for language understanding. In: Advances in Neural Information Processing Systems 32, pp 5753-5763
48. He P, Liu X, Gao J, Chen W (2021) DeBERTa: decoding-enhanced BERT with disentangled attention. In: International Conference on Learning Representations. https://arxiv.org/abs/2006.03654
49. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 21(140):1-67
50. Goel K, Rajani NF, Vig J, Taschdjian Z, Bansal M, Ré C (2021) Robustness Gym: unifying the NLP evaluation landscape. In: Proceedings of NAACL-HLT 2021: Demonstrations, pp 42-55. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-demos.6
51. Yao L, Mao C, Luo Y (2019) Graph convolutional networks for text classification. Proc AAAI Conf Artif Intell 33(01):7370-7377. https://doi.org/10.1609/aaai.v33i01.33017370
52. Koh PW, Sagawa S, Marklund H, Xie SM, Zhang M, Balsubramani A, Hu W, Yasunaga M, Phillips RL, Gao I, Lee T, David E, Stavness I, Guo W, Earnshaw BA, Haque IS, Beery S, Leskovec J, Kundaje A, Pierson E, Levine S, Finn C, Liang P (2021) WILDS: a benchmark of in-the-wild distribution shifts. In: Proceedings of the 38th International Conference on Machine Learning. Proc Mach Learn Res 139:5637-5664. https://proceedings.mlr.press/v139/koh21a.html
53. Beltagy I, Peters ME, Cohan A (2020) Longformer: the long-document transformer. arXiv:2004.05150. https://doi.org/10.48550/arXiv.2004.05150
54. Jiang Z, Yang MY, Tsirlin M, Tang R, Dai Y, Lin J (2023) “Low-Resource” text classification: a parameter-free classification method with compressors. In: Findings of the Association for Computational Linguistics: ACL 2023, pp 6810-6828. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.426
55. Zaheer M, Guruganesh G, Dubey KA, Ainslie J, Alberti C, Ontanon S, Pham P, Ravula A, Wang Q, Yang L, Ahmed A (2020) Big Bird: transformers for longer sequences. In: Advances in Neural Information Processing Systems 33, pp 17283-17297
56. Bulla L, Gangemi A, Mongiovì M (2023) Towards distribution-shift robust text classification of emotional content. In: Findings of the Association for Computational Linguistics: ACL 2023, pp 8256-8268. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.524
57. Li Z, Zhu H, Lu Z, Yin M (2023) Synthetic data generation with large language models for text classification: potential and limitations. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp 10443-10461. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.647
58. Siino M, Tinnirello I, La Cascia M (2024) Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. Inf Syst 121:102342. https://doi.org/10.1016/j.is.2023.102342
59. Casanueva I, Temčinas T, Gerz D, Henderson M, Vulić I (2020) Efficient intent detection with dual sentence encoders. In: Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, pp 38-45. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.nlp4convai-1.5
60. Patwa P, Filice S, Chen Z, Castellucci G, Rokhlenko O, Malmasi S (2024) Enhancing low-resource LLMs classification with PEFT and synthetic data. In: Proceedings of LREC-COLING 2024, pp 6017-6023. ELRA and ICCL. https://aclanthology.org/2024.lrec-main.533/
61. Chen PH, Chen YN (2024) Efficient unseen language adaptation for multilingual pre-trained language models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp 18983-18994. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.1057

62. Li Y, Bonatti R, Abdali S, Wagle J, Koishida K (2024) Data generation using large language models for text classification: an empirical case study. In: ICML 2024 Workshop on Data-centric Machine Learning Research (DMLR): Datasets for Foundation Models. arXiv:2407.12813. https://doi.org/10.48550/arXiv.2407.12813
63. Bugueño M, de Melo G (2023) Connecting the dots: what graph-based text representations work best for text classification using graph neural networks? In: Findings of the Association for Computational Linguistics: EMNLP 2023, pp 8943-8960. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.600
64. Li L, Sleem L, Gentile N, Nichil G, State R (2025) Small language models in the real world: insights from industrial text classification. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pp 971-982. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-industry.68
65. Chen J, Xiao S, Zhang P, Luo K, Lian D, Liu Z (2024) M3-Embedding: multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In: Findings of the Association for Computational Linguistics: ACL 2024, pp 2318-2335. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.137
66. M Y, Singh V, Maheshwari A, Krishna A, Ramakrishnan G (2025) ARISE: iterative rule induction and synthetic data generation for text classification. In: Findings of the Association for Computational Linguistics: NAACL 2025, pp 6434-6449. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-naacl.359
67. Estevanell-Valladares EL, Estévez-Velarde S, Gutiérrez Y, Montoyo A, Mitkov R (2025) XAutoLM: efficient fine-tuning of language models via meta-learning and AutoML. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp 7863-7880. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.399
68. Nguyen QA, Tomeh N, Lebbah M, Charnois T, Azzag H (2025) Leveraging text-to-text transformers as classifier chain for few-shot multi-label classification. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp 26941-26950. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1368
69. Qin C, Chen X, Li J, Zhu H (2026) BOLT: benchmarking open-world learning for text classification. In: Findings of the Association for Computational Linguistics: ACL 2026, pp 13624-13658. Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.findings-acl.667
70. Cunha W, Rocha L, Gonçalves MA (2025) A thorough benchmark of automatic text classification: From traditional approaches to large language models. arXiv:2504.01930. https://doi.org/10.48550/arXiv.2504.01930
71. Siino M, Tinnirello I, La Cascia M (2025) The text classification pipeline: starting shallow going deeper. arXiv:2501.00174. https://doi.org/10.48550/arXiv.2501.00174
72. Zhou C, Dong J, Huang X, Liu Z, Zhou K, Xu Z (2024) QUEST: efficient extreme multi-label text classification with large language models on commodity hardware. In: Findings of the Association for Computational Linguistics: EMNLP 2024, pp 3929-3940. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.226
73. Bertalis N, Granse P, Gül F, Hauss F, Menkel L, Schüler D, Speier T, Galke L, Scherp A (2024) Your extreme multi-label classifier is secretly a hierarchical text classifier for free. arXiv:2411.13687. https://doi.org/10.48550/arXiv.2411.13687
74. Le Bail M, Dentan J, Buscaldi D, Vanier S (2025) Unveiling decision-making in LLMs for text classification: extraction of influential and interpretable concepts with sparse autoencoders. arXiv:2506.23951. https://doi.org/10.48550/arXiv.2506.23951
75. Sung M, Gung J, Mansimov E, Pappas N, Shu R, Romeo S, Zhang Y, Castelli V (2023) Pre-training intent-aware encoders for zero- and few-shot intent classification. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp 10433-10442. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.646
76. Yehudai A, Vetzler M, Mass Y, Lazar K, Cohen D, Carmeli B (2023) QAID: Question Answering Inspired Few-Shot Intent Detection. arXiv:2303.01593. https://doi.org/10.48550/arXiv.2303.01593
77. Dai X, Chalkidis I, Darkner S, Elliott D (2022) Revisiting transformer-based models for long document classification. In: Findings of the Association for Computational Linguistics: EMNLP 2022, pp 7212-7230. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.534
78. Chung HW, Hou L, Longpre S, et al (2024) Scaling instruction-finetuned language models. J Mach Learn Res 25(70):1-53
79. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53-65. https://doi.org/10.1016/0377-0427(87)90125-7
80. Cox DR (1972) Regression models and life-tables. J R Stat Soc Series B Stat Methodol 34(2):187-202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x
81. Lhoest Q, Villanova del Moral A, Jernite Y, et al (2021) Datasets: a community library for natural language processing. In: Proceedings of EMNLP 2021: System Demonstrations, pp 175-184. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-demo.21

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.pdf
- ESM2.zip