

Representation choice and training-boundary composition in document stream segmentation

Dmitriy Ulybin*

*TaxDome, New York, United States*ORCID: [0009-0007-1073-3266](https://orcid.org/0009-0007-1073-3266)* Corresponding author: Dmitriy Ulybin, dv.ulybin@gmail.com

OPEN ACCESS

Citation:

Dmitriy Ulybin (2026). Representation choice and training-boundary composition in document stream segmentation. *Am. Impact Rev.*
[10.66308/air.e2026086](https://doi.org/10.66308/air.e2026086)

Received: March 3, 2026

Accepted: April 8, 2026

Published: April 13, 2026

DOI:

[10.66308/air.e2026086](https://doi.org/10.66308/air.e2026086)

ISSN: 3071-124X

Copyright:

© 2026 Dmitriy Ulybin. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0).

Abstract

This study evaluates representation choice and training-boundary composition for document stream segmentation. On Dutch OpenPSS LONG and SHORT, four document orders were compared across five training seeds while holding document exposure, boundary prevalence and pair multiplicity fixed. Term frequency–inverse document frequency (TF–IDF) features outperformed frozen multilingual MiniLM in all eight corresponding comparisons by 5.22–7.26 percentage points of boundary F1. Native ordering gave the highest LONG means. Similarity-based construction yielded a difference of –0.55 points versus random ordering in the prespecified LONG MiniLM comparison (95% bootstrap confidence interval [–1.19, +1.71]). A separate extension compared four procedures on LONG and English TABME++, using 10,000 training pairs, 3,000 validation pairs and three seeds per corpus. Selective adaptation updated the final four encoder layers for two epochs. On TABME++, it achieved 78.42% boundary F1 versus 73.72% for the matched frozen control, a difference of +4.71 points (conditional 95% cluster interval [+3.20, +6.29]). TF–IDF and frozen MiniLM with stochastic gradient descent achieved 86.43% and 85.23%, respectively. Exact document recovery and local optimization time complemented boundary scoring. The findings support comparing lexical and adapted representations under explicit training budgets and retaining observed document order as a practical baseline.

Keywords: document stream segmentation, selective fine-tuning, TF–IDF, sentence embeddings, training data composition, duplicate screening, model evaluation

1. Introduction

Document stream segmentation recovers individual documents from an ordered sequence of pages. It supports downstream classification, extraction and retrieval: a missed transition can merge unrelated records, whereas an unnecessary cut can separate a form from its continuation. OpenPSS provides public administrative streams for studying these errors [1]. TABME constructs business-document streams from archived documents [2], and TABME++ extends this setting with improved OCR and language-model experiments [3]. Together, these resources make document segmentation accessible to reproducible experimental research.

Two choices shape a text-based segmentation pipeline. The first is the representation of each page: lexical features, frozen pretrained embeddings or representations adapted using boundary labels. The second is the construction of training streams. Concatenating the same documents in different orders changes cross-document examples even when every page is retained. These choices can interact with vocabulary, templates

and the distribution of document transitions encountered at evaluation.

Lexical methods warrant a substantive comparison. Wang and Manning showed that simple linear models and appropriate lexical features can be strong text-classification baselines [4]. Their sentiment and topic classification findings motivate a direct comparison of lexical and pretrained representations. Likewise, BERT distinguishes fixed feature extraction from supervised fine-tuning [5]. These are distinct representation strategies with different training requirements.

Training-boundary composition raises a separate hypothesis. Random concatenation can connect topically unrelated documents, making a boundary visible as a change in vocabulary. Selecting more similar neighbors could expose a classifier to transitions that lack that cue. Yet increased similarity may instead shift the training distribution away from real streams. A useful test must separate this intervention from changes in class proportions, repeated exposure or the documents available for learning.

This study compares four controlled training orders on two OpenPSS collections using term frequency–inverse document frequency (TF–IDF) classifiers and frozen multilingual embeddings. A second experiment evaluates selective encoder adaptation on LONG and English TABME++ using common labeled-pair budgets. The study measures boundary and exact-document recovery, paired performance differences and local optimization time. Together, these experiments examine the contribution of training-boundary composition and compare lexical, frozen and adapted representations for the same segmentation task.

2. Related work

Segmentation quality has several levels. Insurance-document research evaluates pages, documents and complete streams while connecting confidence to automation [6]. DocSplit addresses document-packet recognition and splitting across a broader benchmark [7]. Multimodal approaches combine textual and visual evidence [8], and work beyond page classification emphasizes the structural demands of document understanding [9]. These studies motivate evaluating exact document recovery alongside local boundary decisions. Our design examines adjacent-page classification using the supplied OCR.

Synthetic document streams are established in TABME and TABME++. The latter discusses how deduplicating sampled pairs changes class proportions [3]. We extend this line of work through an exposure-controlled comparison in which within-document examples, endpoints, class counts and pair multiplicities remain fixed while cross-document neighbors change. Lexical similarity determines the new neighbors.

Sentence-BERT provides sentence representations suited to similarity calculations [10], and multilingual knowledge distillation extends sentence embeddings across languages [11]. MiniLM compresses pretrained transformer representations through self-attention distillation [12]. A task-specific linear head learns a decision rule over these reusable page features, while supervised fine-tuning updates the representations themselves.

Repeated training also requires careful interpretation. Reimers and Gurevych demonstrate why single neural scores can obscure variation across runs [13]. Dodge and colleagues study initialization and training-example presentation order during language-model fine-tuning [14]. Their ordering intervention rearranges existing examples; ours changes which page pairs exist at document transitions. Reporting selection budgets matters because best-found validation performance depends on the number of trials [15]. More generally, data sampling, initialization and hyperparameter choice contribute distinct sources of benchmark variance [16]. We distinguish training-seed variation from variation in the evaluated source streams.

3. Materials and methods

3.1 Data and content screening

We used the original OpenPSS optical character recognition (OCR) text and boundary labels. LONG contains 63,815 released training pages and 25,676 test pages; SHORT contains 40,715 training pages and 11,462 test pages [1]. LONG comprises original administrative streams. SHORT reconstructs streams by concatenating documents in archive order. Both collections contain Dutch administrative material. Images were verified against the released page records but were not supplied to the classifiers.

The source releases were recorded with URLs, sizes and SHA-256 hashes. Inspection identified training/test streams with identical ordered OCR and labels, including LONG pairs containing 72 and 5,187 pages. Stream-identifier separation alone was consequently inadequate. Avoiding repeated evaluation content is a recognized motivation for deduplication [17]; our procedure used text-shingle resemblance in the tradition of Broder [18] to screen document content before fitting.

Normalization case-folded text and collapsed whitespace while retaining numbers and punctuation. Exact matches required at least 200 normalized characters. Documents of at least 500 characters were also compared using word-trigram sets. MinHash sketches with 128 permutations and a locality-sensitive hashing (LSH) threshold of 0.85 generated candidate near matches; acceptance required exact Jaccard similarity of at least 0.90 and a normalized-length ratio of at least 0.80. Connected components joined exact and accepted near matches.

Twenty percent of original training sources, rounded upward, were assigned to validation with seed 20260927. Cross-split components were resolved with precedence test, validation, then training. Entire documents were removed from lower-priority splits while original test pages were preserved. A removal divided the remaining source into contiguous blocks, preventing invented adjacency across the deleted material. “Native” henceforth means the original relative order within these retained blocks.

The process excluded 7,336 development pages from LONG and 3,702 from SHORT. Table 1 and Figure 1 summarize the retained population. Content screening used OCR alone and was completed before fitting. Six LONG streams also lack leading pages, totaling 107 unavailable page numbers; evaluation covers observed sequences rather than reconstructed prefixes.

Table 1. Retained data following source-level validation assignment and document exclusions. Boundary prevalence excludes the known first page of each block. Blocks produced by exclusions remain associated with their original sources.

Collection / split	Pages	Documents	Source streams	Retained blocks	Boundary rate (%)
LONG / train	43,185	13,010	58	194	29.81
LONG / validation	13,294	3,843	16	31	28.74
LONG / test	25,676	6,347	34	34	24.62
SHORT / train	28,748	4,389	159	204	14.66
SHORT / validation	8,265	1,838	40	44	21.82
SHORT / test	11,462	1,404	108	108	11.41

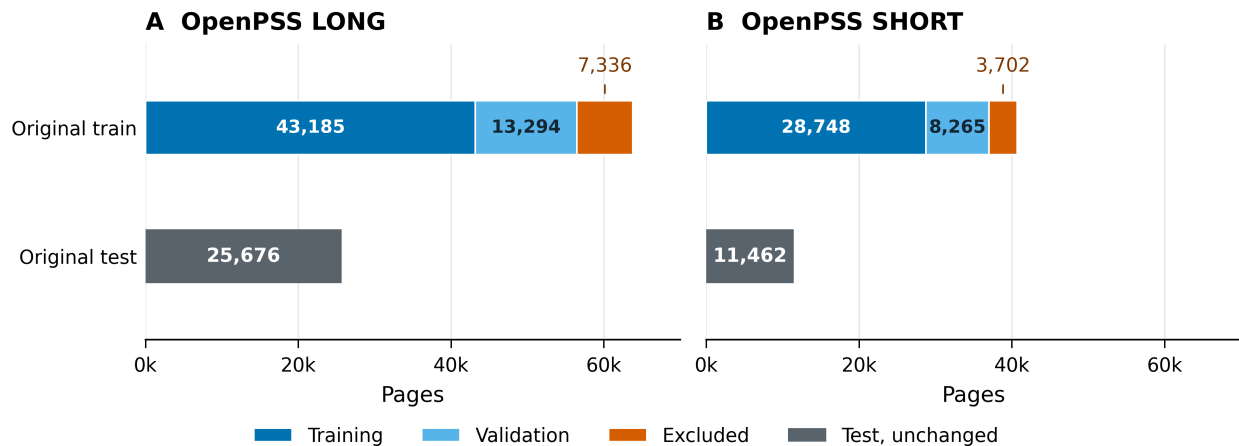


Figure 1. Disposition of page records. The original training partition is divided into retained training, validation and excluded pages. Original test pages are unchanged. Both panels use a common horizontal scale.

3.2 Controlled training-boundary composition

Documents were permuted only within retained training blocks. Internal page sequences and each block's first and last documents were fixed. With M pages, D documents and S blocks, a complete presentation has $D-S$ cross-document pairs and $M-D$ within-document pairs. The constraints thus preserve boundary prevalence as well as document exposure. Blocks with fewer than two internal documents remain unchanged.

Native retains the source order. Random uniformly permutes the internal documents. Similar greedily selects the available internal document whose first page has the highest TF-IDF cosine similarity to the preceding document's final page; ties are randomized. Mixed assigns half, rounded downward, of the non-forced internal placement decisions to this similarity rule and the remainder to uniform random selection. Mixed is therefore a mixture of decisions rather than precisely half of all observed boundaries.

Each seed generates one complete order per condition, repeated twice. This construction was fixed before supervised fitting to match the number and multiplicity of unique page-ID pairs; combining two different complete orders would not provide that match with duplicated Native. Assertions checked identical within-document pair multisets, page exposure, class counts, endpoints and pair multiplicities. Reordering was possible in 141 of 194 LONG blocks and 143 of 204 SHORT blocks. Construction and classifier randomization used seeds 17, 29, 43, 71 and 101.

3.3 Page representations and boundary classifiers

The encoder was sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2, pinned to revision e8f8c211226b894fcb81acc59f3b34ba3efd5f42, with verified released weights [19]. It has 117,653,760 parameters and returns 384-dimensional vectors. Encoder weights remained frozen in the training-order experiment.

Each page received up to 126 content tokens plus two special tokens. Longer pages retained the first 63 and last 63 content tokens. Attention-masked mean pooling, including special tokens, followed by L2 normalization produced the page vector. The lexical comparator used 10,000 unigram/bigram TF-IDF features, minimum document frequency 2 and sublinear term frequency. Its vocabulary and inverse document frequencies were fitted on retained training pages only. It consumed decoded text from the same selected tokens, giving both representations the same selected textual content.

For vectors u and v of adjacent pages, the classifier input concatenated u , v , their absolute difference, elementwise product and cosine similarity, plus $\log(1 + \text{raw OCR character count})/10$ for each page. Feature dimensions were 1,539 for MiniLM and 40,003 for TF-IDF. The character counts supply identical full-page length information to both methods. Filenames, document identifiers, source labels and external page indices were excluded.

Both representations used scikit-learn logistic-loss stochastic gradient descent (SGD) classifiers [20], L2 regularization with alpha 0.0001, optimal learning-rate scheduling and 30 shuffled passes. There was no class weighting, early stopping or hyperparameter search. The duplicated tables produce 60 presentations of each unique pair. The fixed classifier settings and optimization counts make each representation comparison reproducible.

For each fit, validation boundary F1 selected a threshold from 0.05 to 0.95 in increments of 0.01; ties favored the value nearest 0.50 and then the smaller value. Four Native pilot fits checked infrastructure and validation behavior and were retained with the unchanged configuration. All 80 fitted classifiers and selected thresholds were saved before test scoring. Computation used an Apple M4 Pro computer with 24 GiB memory, MPS encoder inference and CPU classifier fitting.

3.4 Outcomes and inference

Boundary precision, recall and F1 pool confusion counts across test streams within a seed, excluding each stream's forced first-page start. The reported mean is the arithmetic average of five seed-level metrics. Exact document precision, recall and F1 require both endpoints of a predicted interval to match one labeled document in the same observed stream, consistent with prior exact-match evaluation [3, 6]. Missed-boundary rate uses true boundaries as its denominator; extra-cut rate uses true within-document adjacencies. Reporting both rates separates missed transitions from unnecessary cuts.

The prespecified primary contrast is Similar minus Random boundary F1 for frozen MiniLM on LONG. Bootstrap resampling [21] and paired evaluation [22] motivate uncertainty estimation using shared samples. Detected duplicate-document links joined test sources into 26 LONG and 103 SHORT clusters. For each of 2,000 draws, five seed indices and the original number of clusters were sampled with replacement, using the same selections across compared procedures. Counts were pooled within each sampled seed before calculating F1, and the seed-level values were averaged. Reported 95% confidence intervals (CIs) are percentile intervals for these paired differences.

Work on crossed data arrays motivates accounting for dependence across shared observations and factors [23]. Our paired resampling retains seed and source-cluster structure and provides approximate intervals conditional on the development split. Secondary comparisons are descriptive and unadjusted. Prespecified sensitivity checks used threshold 0.50 and omitted the five largest test streams without refitting.

After inspecting the original results, additional analyses compared TF-IDF with MiniLM and Native with Random using the same paired draws. These analyses did not modify models or thresholds and remain post hoc and descriptive. Leave-one-cluster-out analysis removed each complete duplicate-linked cluster and recomputed the paired mean difference. Its range of point estimates diagnoses sensitivity to individual clusters. We also measured the distribution of test pages and true boundaries across sources and clusters.

3.5 Selective encoder adaptation on matched training samples

The adaptation experiment compares four procedures on LONG Native and English TABME++ [3]: TF-IDF with SGD, frozen MiniLM with SGD, frozen MiniLM with an AdamW-trained linear head, and MiniLM with

its final four transformer layers updated alongside the same head. The first two retain the classifier settings in Section 3.3; the third provides the matched control for encoder adaptation. Both corpora use the pinned encoder and page-input policy described above.

TABME++ retains the released document-level splits. Applying the same content-overlap rules excluded 23 training documents containing 127 pages. The retained training, validation and test partitions contain 40,269, 2,238 and 2,238 documents, respectively, comprising 110,005, 6,089 and 6,237 pages. Within each split, lexicographically sorted document IDs were permuted using one NumPy generator with seed 20260928, applied sequentially to training, validation and test. Documents were assembled into packets of ten, retaining incomplete final packets and each document's original page order. This yielded 4,027 training packets and 224 packets in each evaluation partition. Each packet's first page is a known start, and page pairs never cross packets.

Each corpus supplies a fixed sample of 10,000 unique training pairs and 3,000 validation pairs, shared by all four procedures and seeds 17, 29 and 43. Sampling is without replacement within boundary classes, using proportional allocation with largest remainders and fixed sampling seeds 20260928 and 20260929. LONG contains 2,981 training and 862 validation boundaries; TABME++ contains 3,420 and 1,030. TF-IDF vocabulary and inverse document frequencies are fitted only on distinct pages appearing in the selected training pairs. Evaluation uses the complete unchanged test populations: 25,676 LONG pages and 6,237 TABME++ pages.

Selective adaptation updates 7,097,856 encoder parameters in the final four transformer layers; embeddings and the first eight layers remain fixed. Neural fits use a 1,539-input linear head, binary cross-entropy without class weights, two epochs, batches of 16 and float32 computation. Frozen and adapted variants share initial head weights, minibatch permutations, update counts and nominal head optimization. AdamW [24] uses learning rates of 0.001 for the head and 0.00002 for updated encoder parameters, betas (0.9, 0.999), and epsilon 10^{-8} . Encoder weight decay is 0.01, excluding biases and LayerNorm parameters; head decay is zero. Both rates use 10% linear warmup followed by linear decay to zero. Dropout is disabled, and a global gradient norm of 1 clips all trainable parameters. Optimization traces and parameter checks document training behavior, consistent with the attention to fine-tuning stability in prior work [25]. SGD procedures retain 60 presentations per selected pair, compared with two for the neural procedures.

After each neural epoch, validation boundary F1 selects the threshold on the grid defined in Section 3.3. The highest validation F1 determines the checkpoint, with the earlier epoch retained on a tie. SGD thresholds use the same validation sample and grid. All 24 fits and selection decisions were finalized before test evaluation.

The primary extension contrast is selective adaptation minus its matched frozen control on TABME++; LONG is exploratory because its test results were examined in the original experiment. Metrics pool test counts within each seed and then average the three seed-level values. For 2,000 paired bootstrap draws, duplicate-linked test clusters are resampled while all three fitted seeds remain fixed. There are 26 LONG and 224 TABME++ clusters. Percentile 95% intervals describe conditional sampling uncertainty; seed-specific contrasts, threshold-0.50 results and leave-one-cluster-out estimates accompany the main comparison.

3.6 Reproducibility

The original protocol was recorded locally before supervised fitting. The paired representation and Native-order analyses were added after the original results; the selective-adaptation extension was specified separately before its new evaluation. The research package retains source and model hashes, development assignments, exclusions, construction orders, fit parameters, seed-level predictions and figure data. Independent scoring code verified boundary F1 and reconstructed exact documents from actual page identifiers. The saved manifests

connect each reported estimate to its inputs and predictions.

4. Results

4.1 Representation choice and complete-document recovery

TF-IDF produced higher mean boundary F1 than frozen MiniLM for every corresponding order and collection (Table 2; Figure 2). LONG Native achieved 84.75% with TF-IDF and 78.82% with MiniLM. The corresponding SHORT values were 55.02% and 48.64%. The difference is consistent across the examined procedures, although the absolute quality differs substantially between collections.

Table 2. Mean test boundary F1 / exact document F1 across five seeds, expressed as percentages. MiniLM encoder weights are frozen. All four planned training orders are shown; seed-level scores and standard deviations are retained in the research package.

Training order	LONG: MiniLM	LONG: TF-IDF	SHORT: MiniLM	SHORT: TF-IDF
Native	78.82 / 58.94	84.75 / 70.09	48.64 / 28.13	55.02 / 36.57
Random	78.34 / 58.23	83.70 / 67.69	49.12 / 27.99	55.31 / 36.29
Similar	77.80 / 57.21	83.02 / 67.24	46.25 / 25.83	52.42 / 34.23
Mixed	78.57 / 58.59	84.31 / 69.40	47.28 / 26.63	54.54 / 36.56

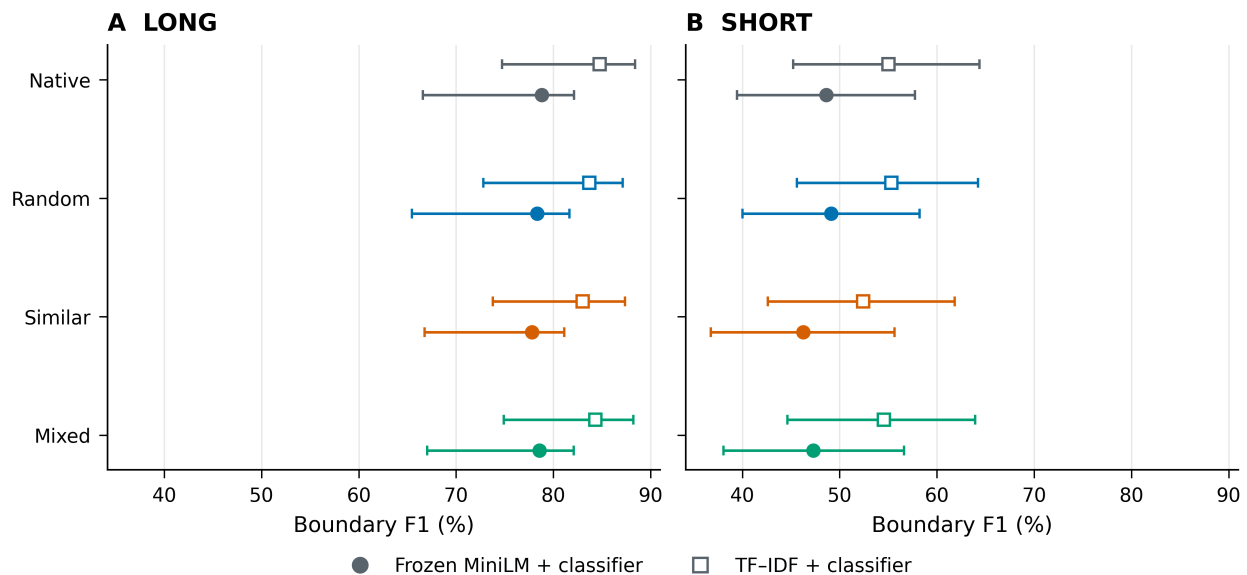


Figure 2. Mean boundary F1 with 95% seed/cluster percentile bootstrap intervals. Filled circles denote frozen MiniLM and open squares denote TF-IDF. Absolute intervals do not replace paired contrasts between procedures.

Exact document F1 was lower than boundary F1 in every condition. LONG Native yielded 58.94% document F1 for MiniLM and 70.09% for TF-IDF, despite boundary F1 values of 78.82% and 84.75%. SHORT document F1 ranged from 25.83% to 36.57%. Local decisions consequently left substantial room for improvement in full-document recovery. The two outcomes describe complementary aspects of the same segmentation output.

4.2 Training-boundary composition

Similarity-based construction raised mean training-boundary cosine similarity from 0.0624 under Random to 0.2036 under Similar on LONG. Native and Mixed yielded 0.0862 and 0.1399. On SHORT, the corresponding means were 0.0915, 0.2324, 0.1198 and 0.1650. Exposure and class-balance assertions passed throughout. The construction increased lexical continuity across document transitions while preserving the controlled exposure quantities.

The primary outcome did not establish a benefit from that change. LONG MiniLM boundary F1 was 77.80% for Similar and 78.34% for Random: -0.55 percentage points, with a paired 95% interval of $[-1.19, +1.71]$ (Figure 3). The secondary LONG TF-IDF contrast was -0.68 points $[-1.34, +2.30]$. On SHORT, Similar-minus-Random differences were -2.87 points $[-4.64, -1.11]$ for MiniLM and -2.89 points $[-4.37, -1.40]$ for TF-IDF.

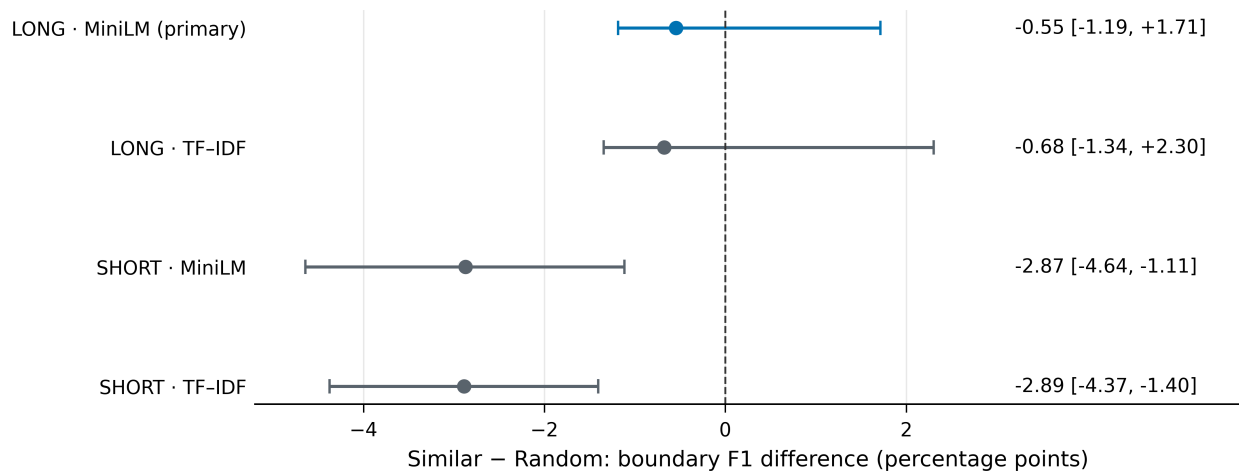


Figure 3. Similar-minus-Random differences in mean boundary F1, in percentage points, with 95% percentile intervals from 2,000 paired seed/cluster draws. LONG MiniLM is the prespecified primary comparison; the others are descriptive. The vertical reference marks zero difference.

For LONG MiniLM, Similar missed 23.12% of true boundaries versus 21.37% under Random, while extra-cut rates were 6.78% and 7.22%, respectively. Thus, fewer unnecessary cuts coincided with more missed transitions. At threshold 0.50, Similar-minus-Random boundary F1 remained negative in all four corpus/representation combinations. Removing the five largest test streams reduced the LONG MiniLM difference to approximately -0.05 points and changed the LONG TF-IDF difference to approximately $+0.24$ points; SHORT differences remained negative.

4.3 Representation and native-order differences

Paired TF-IDF-minus-MiniLM boundary F1 differences ranged from $+5.22$ to $+5.93$ percentage points across LONG orders and from $+6.17$ to $+7.26$ points across SHORT orders. All eight descriptive intervals lay above zero. LONG Native differed by $+5.93$ points $[5.22, 9.17]$, and SHORT Native by $+6.38$ points $[3.20, 10.04]$. Interval widths ranged from 3.60 to 3.95 points on LONG and 6.39 to 7.02 on SHORT, showing appreciable uncertainty around their magnitudes.

Removing any single source cluster preserved the positive representation difference in all eight conditions. For Native, leave-one-cluster-out differences ranged from $+5.71$ to $+6.95$ points on LONG and from $+5.56$ to $+7.15$ on SHORT. Exact document F1 differences also favored TF-IDF: $+9.46$ to $+11.14$ points on LONG and

+8.31 to +9.93 on SHORT. The observed representation difference therefore extends from local boundary classification to complete-document recovery.

Native ordering had the highest observed LONG boundary means. Its post hoc advantage over Random was +0.48 points [0.09, 1.48] for MiniLM and +1.05 points [0.62, 2.95] for TF-IDF. Both directions persisted after removing any one cluster. SHORT did not show this pattern: Native-minus-Random differences were -0.48 points [-1.30, +0.43] and -0.29 points [-1.12, +0.61], respectively. Retaining observed adjacency was useful on LONG, whereas Random produced the stronger boundary means on SHORT.

4.4 Source concentration and robustness

The LONG test contains 34 sources but only 26 detected duplicate-linked clusters. Its largest cluster joins eight sources and contains 18,362 pages - 71.51% of test pages - and 62.22% of true boundaries. The five largest individual streams all belong to this cluster. SHORT is less concentrated: its largest cluster contains three sources and 13.83% of pages.

The LONG primary MiniLM contrast remained negative after every single-cluster removal, ranging from -0.76 to -0.08 points. LONG TF-IDF Similar-minus-Random changed sign after removal of one cluster, with a range from -0.84 to +0.48 points. SHORT ordering differences remained negative throughout. These results distinguish the comparatively stable representation ranking from a smaller, more distribution-sensitive ordering effect.

4.5 Selective adaptation on LONG and TABME++

The four procedures used the same 10,000 selected training pairs and 3,000 validation pairs within each corpus. Table 3 and Figure 4 report test boundary F1 and exact document F1.

Table 3. Test performance and local optimization time using shared training pairs. Values are means across three seeds. Optimization time measures the training loop only: 60 presentations per pair for SGD and two for the neural procedures.

Neural timing includes forward/backward computation and batch feature assembly. Timings exclude OCR, offline preprocessing, validation, model loading and test prediction. Measurements used the same computer with possible concurrent workloads.

Corpus	Procedure	Boundary F1 (%)	Exact document F1 (%)	Optimization time (s)
LONG	TF-IDF + SGD	82.89	67.07	0.5
LONG	Frozen MiniLM + SGD	77.99	57.97	2.3
LONG	Matched frozen MiniLM	65.18	36.25	8.3
LONG	Selective MiniLM (last 4)	81.67	65.44	510.2
TABME++	TF-IDF + SGD	86.43	71.33	0.3
TABME++	Frozen MiniLM + SGD	85.23	68.17	1.4
TABME++	Matched frozen MiniLM	73.72	50.84	7.3
TABME++	Selective MiniLM (last 4)	78.42	56.44	391.1

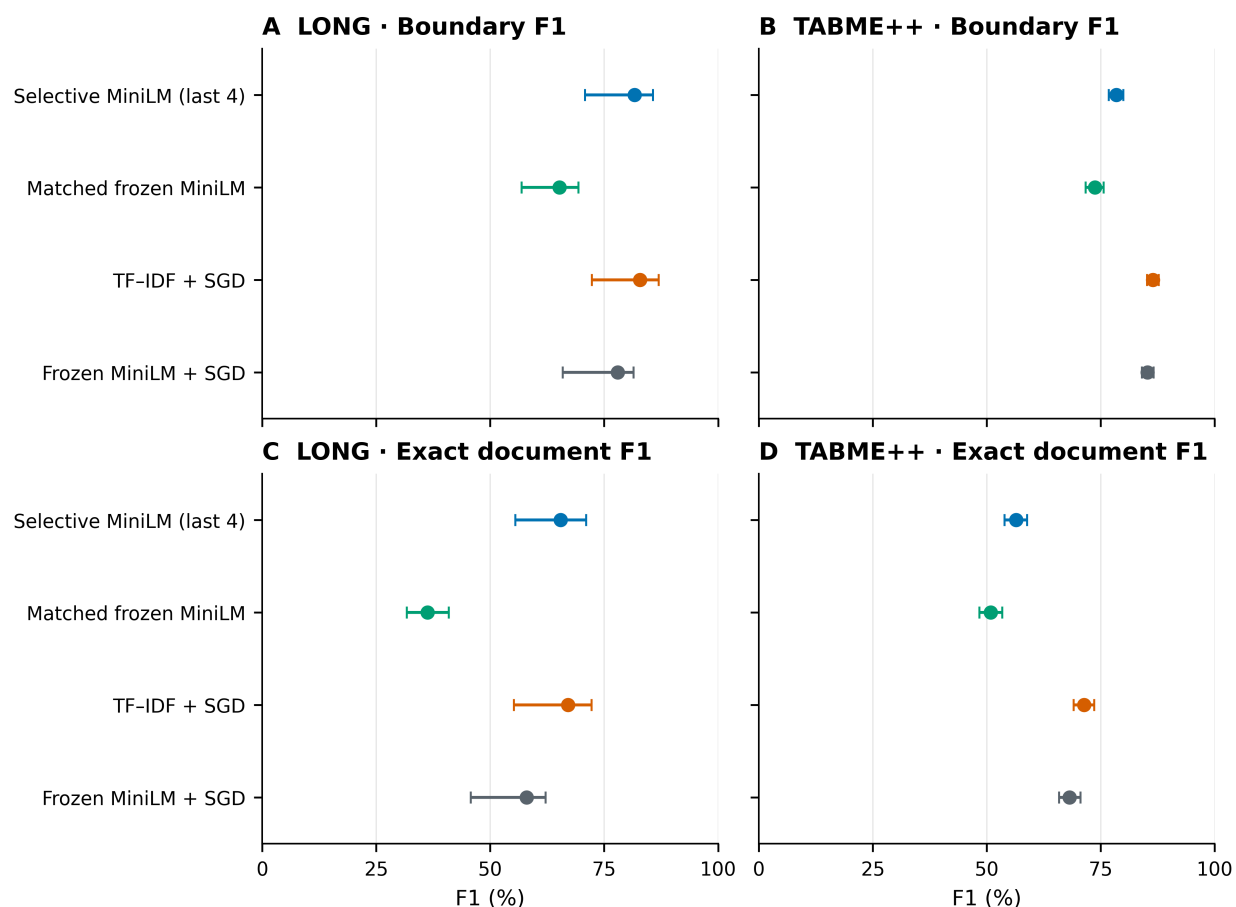


Figure 4. Boundary and exact document F1 for the four procedures. Points are means across three seeds; lines show 95% percentile intervals from 2,000 paired resamples of test source clusters, holding the trained seeds fixed. Selective MiniLM updates the last four encoder layers. All procedures share the selected training and validation pairs. TABME++ comprises fixed, document-disjoint synthetic packets.

On TABME++, selective adaptation achieved 78.42% boundary F1, compared with 73.72% for the matched frozen control. The prespecified extension contrast was +4.71 percentage points (95% CI +3.20 to +6.29). The corresponding LONG difference was +16.49 percentage points (95% CI +13.46 to +17.46) (Figure 5).

On LONG, the three seed-specific differences ranged from +16.27 to +16.71 points; removing individual clusters yielded point differences from +15.68 to +16.77 points. On TABME++, the three seed-specific differences ranged from +4.54 to +4.92 points; removing individual clusters yielded point differences from +4.43 to +4.83 points.

Selective adaptation versus TF-IDF gave the following descriptive boundary-F1 differences: LONG: -1.22 percentage points (95% CI -3.92 to +0.92); TABME++: -8.01 percentage points (95% CI -9.51 to -6.45). The SGD and neural comparisons therefore describe alternative training procedures under a common labeled-pair budget.

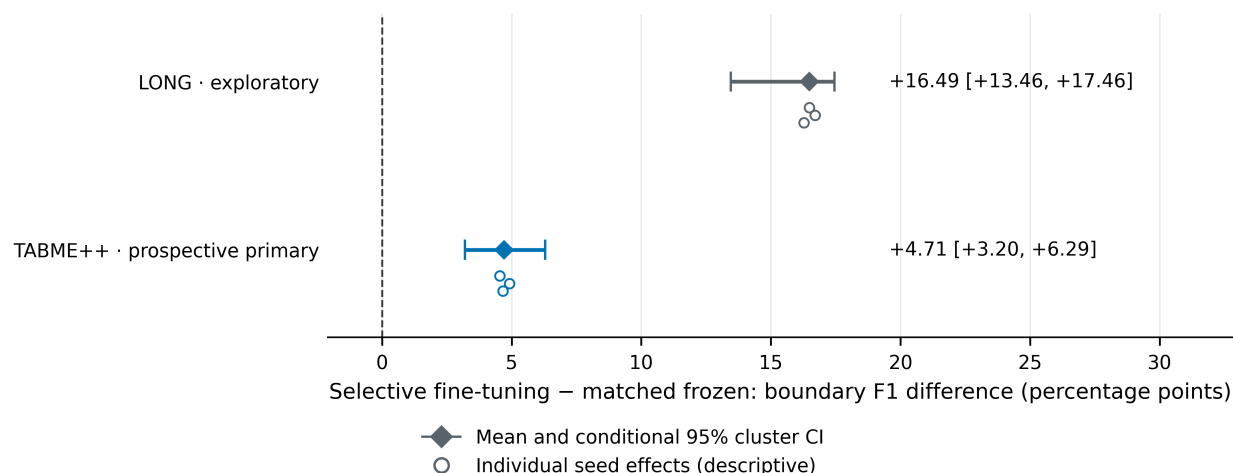


Figure 5. Differences in boundary F1 between selective fine-tuning and the matched frozen control. Diamonds show the mean across three seeds; lines show conditional 95% paired-cluster intervals. Hollow circles show individual seed differences. TABME++ is the prespecified extension comparison; LONG is exploratory. The dashed line marks zero.

5. Discussion

Representation choice produced the clearest observed difference in the original controlled experiment. Across both collections and all training orders, TF-IDF supported higher boundary and exact-document F1 than the frozen MiniLM procedure. Paired analysis and single-cluster deletion preserved the boundary-level direction. This supports including a trained lexical classifier as a serious candidate in document pipelines alongside reusable pretrained features.

Representation choice includes both feature extraction and the task-specific decision rule. Lexical features preserve explicit word patterns, whereas the selected encoder provides dense pretrained features. Earlier lexical and BERT research [4, 5] motivates evaluating these strategies together with supervised adaptation. The substantial gap between boundary and document F1 also suggests evaluating richer context and structural consistency, rather than optimizing only isolated page-pair decisions.

The adaptation experiment separates two practical decisions: whether to update the encoder, and how to train its boundary classifier. With the neural head recipe held fixed, selective adaptation changed mean boundary F1 by +16.49 points on LONG and +4.71 points on TABME++. The separately trained frozen SGD comparator and TF-IDF procedure provide the broader engineering comparison in Table 3. On TABME++, frozen MiniLM with SGD achieved 85.23%, compared with 78.42% after selective adaptation. The adaptation gain is therefore specific to the matched two-epoch neural training recipe, and classifier optimization remains consequential for representation selection. These comparisons evaluate complete training procedures on common labeled pairs; the shared-data design makes their observed differences relevant to representation selection.

Training-boundary construction produced smaller and corpus-dependent effects. The similarity intervention worked mechanically but did not establish a primary predictive advantage. Increasing lexical continuity across boundaries may weaken a cue useful on native evaluation streams, or the fixed representations may fail to express the distinctions required by those examples. These explanations motivate subsequent experiments on the cues used by boundary classifiers. Native’s stronger LONG performance also makes retained source order a useful practical comparator.

The contrast between LONG and SHORT underlines the importance of the evaluation distribution. SHORT uses reconstructed archive order and has markedly different validation and test boundary prevalences. Its lower absolute quality and different ordering ranking make corpus-level evaluation consequential. LONG's source concentration further explains why cluster-based uncertainty and influence diagnostics add information beyond pooled page counts. The paired representation analyses provide a coherent descriptive account of these evaluated sources.

Local optimization time provided a second comparison axis. TF-IDF classifier fitting averaged 0.5 s on LONG and 0.3 s on TABME++; selective encoder adaptation averaged 510.2 s and 391.1 s, respectively. These measurements cover optimization after offline input preparation on the same computer; neural timings include forward/backward computation and batch feature assembly. Parameter count, runtime, memory and operation counts can differ as efficiency measures [26]. A deployment assessment additionally includes OCR, tokenization, feature preparation, model loading and inference under the intended workload. The measured quality and training time provide concrete inputs to that assessment.

The evaluation covers fixed development splits in two Dutch administrative collections and one English constructed-packet population. The shared text budget truncates approximately 82% of OpenPSS pages, and the models use OCR without visual page context. LONG estimates reflect a concentrated source distribution, while duplicate screening identifies only detectable text overlap. The adaptation results characterize the specified two-epoch, last-four-layer training recipe; its intervals are conditional on the three fitted models.

6. Conclusions

TF-IDF exceeded frozen multilingual MiniLM in all eight corresponding comparisons in the OpenPSS training-order experiment. Native ordering produced the highest observed LONG means, while similarity-based boundary construction did not establish an advantage in the original primary comparison. In the extension using common training pairs, TF-IDF with SGD had the highest mean boundary F1 on LONG (82.89%) and TABME++ (86.43%). Selective adaptation differed from its matched frozen control by +4.71 points on TABME++. The combined results support evaluating representation, classifier training and document order together, with lexical baselines, exact-document recovery and clearly scoped computational measurements.

Declarations

Author contributions. Dmitriy Ulybin: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review and editing, Visualization, and Project administration.

Funding. This research received no external funding.

Conflicts of interest. Dmitriy Ulybin is employed by TaxDome. The author declares no competing interests related to this study.

Data availability. The original datasets and pretrained model are available from the cited sources. The local research package contains source manifests, code, fixed protocols, exclusions, saved predictions and figure data.

Ethics statement. This computational study used previously released document datasets. It did not recruit participants or access private TaxDome client records.

References

- [1] R. van Heusden, J. Kamps, and M. Marx, “OpenPSS: An Open Page Stream Segmentation Benchmark,” *Linking Theory and Practice of Digital Libraries*, Lecture Notes in Computer Science, vol. 15177, pp. 413–429, 2024. doi:10.1007/978-3-031-72437-4_24.
- [2] T. Mungmeeprued, Y. Ma, N. Mehta, and A. Lipani, “Tab this Folder of Documents: Page Stream Segmentation of Business Documents,” *Proceedings of the 22nd ACM Symposium on Document Engineering*, pp. 1–10, 2022. doi:10.1145/3558100.3563852.
- [3] H. Heidenreich, R. Dalvi, R. Mukku, N. Verma, and N. Pičuljan, “Large Language Models for Page Stream Segmentation,” arXiv:2408.11981v1, 2024, preprint. doi:10.48550/arXiv.2408.11981.
- [4] S. I. Wang and C. D. Manning, “Baselines and Bigrams: Simple, Good Sentiment and Topic Classification,” *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 90–94, 2012. [Online]. Available: <https://aclanthology.org/P12-2018/>.
- [5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. doi:10.18653/v1/N19-1423.
- [6] H. Heidenreich, R. Dalvi, N. Verma, and Y. Getachew, “Page Stream Segmentation with LLMs: Challenges and Applications in Insurance Document Automation,” *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pp. 305–317, 2025. [Online]. Available: <https://aclanthology.org/2025.coling-industry.26/>.
- [7] M. M. Islam, M. S. Salekin, N. Balakrishnan, V. C. Bishop III, N. Jain, S. Romo, et al., “DocSplit: A Comprehensive Benchmark Dataset and Evaluation Approach for Document Packet Recognition and Splitting,” arXiv:2602.15958v1, 2026, preprint. doi:10.48550/arXiv.2602.15958.
- [8] G. Wiedemann and G. Heyer, “Multi-modal Page Stream Segmentation with Convolutional Neural Networks,” *Language Resources and Evaluation*, vol. 55, no. 1, pp. 127–150, 2021. doi:10.1007/s10579-019-09476-2.
- [9] J. Van Landeghem, S. Biswas, M. Blaschko, and M.-F. Moens, “Beyond Document Page Classification: Design, Datasets, and Challenges,” *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2950–2960, 2024. doi:10.1109/WACV57701.2024.00294.
- [10] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019. doi:10.18653/v1/D19-1410.
- [11] N. Reimers and I. Gurevych, “Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation,” *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4512–4525, 2020. doi:10.18653/v1/2020.emnlp-main.365.
- [12] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, “MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 5776–5788, 2020. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.

- [13] N. Reimers and I. Gurevych, “Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging,” *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 338–348, 2017. doi:10.18653/v1/D17-1035.
- [14] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith, “Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping,” arXiv:2002.06305v1, 2020, preprint. doi:10.48550/arXiv.2002.06305.
- [15] J. Dodge, S. Gururangan, D. Card, R. Schwartz, and N. A. Smith, “Show Your Work: Improved Reporting of Experimental Results,” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2185–2194, 2019. doi:10.18653/v1/D19-1224.
- [16] X. Bouthillier, P. Delaunay, M. Bronzi, A. Trofimov, B. Nichyporuk, J. Szeto, et al., “Accounting for Variance in Machine Learning Benchmarks,” *Proceedings of Machine Learning and Systems*, vol. 3, pp. 747–769, 2021. [Online]. Available: https://proceedings.mlsys.org/paper_files/paper/2021/hash/0184b0cd3cfb185989f858a1d9f5c1eb-Abstract.html.
- [17] K. Lee, D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, et al., “Deduplicating Training Data Makes Language Models Better,” *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, 2022. doi:10.18653/v1/2022.acl-long.577.
- [18] A. Z. Broder, “On the Resemblance and Containment of Documents,” *Proceedings. Compression and Complexity of SEQUENCES 1997*, pp. 21–29, 1997. doi:10.1109/SEQUEN.1997.666900.
- [19] Sentence Transformers, “paraphrase-multilingual-MiniLM-L12-v2,” model card and weights, revision e8f8c211226b894fcb81acc59f3b34ba3efd5f42, 2026. [Online]. Available: <https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2/tree/e8f8c211226b894fcb81acc59f3b34ba3efd5f42>.
- [20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011. [Online]. Available: <https://jmlr.org/papers/v12/pedregosa11a.html>.
- [21] B. Efron, “Bootstrap Methods: Another Look at the Jackknife,” *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979. doi:10.1214/aos/1176344552.
- [22] R. Dror, G. Baumer, S. Shlomov, and R. Reichart, “The Hitchhiker’s Guide to Testing Statistical Significance in Natural Language Processing,” *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1383–1392, 2018. doi:10.18653/v1/P18-1128.
- [23] A. B. Owen and D. Eckles, “Bootstrapping Data Arrays of Arbitrary Order,” *The Annals of Applied Statistics*, vol. 6, no. 3, pp. 895–927, 2012. doi:10.1214/12-AOAS547.
- [24] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” *International Conference on Learning Representations*, 2019. [Online]. Available: <https://openreview.net/forum?id=Bkg6RiCqY7>.
- [25] M. Mosbach, M. Andriushchenko, and D. Klakow, “On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines,” *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=nzpLWnVAyah>.

- [26] M. Dehghani, A. Arnab, L. Beyer, A. Vaswani, and Y. Tay, “The Efficiency Misnomer,” *International Conference on Learning Representations*, 2022. [Online]. Available: <https://openreview.net/forum?id=iuLEMLYh1uR>.