Does Locality Cost in Polish Medical Text Classification? Duplicate-Aware Evaluation of Federated Learning

Daniel Cieślak
Gdańsk University of Technology
Multimedia Systems Department
daniel.cieslak@pg.edu.pl

Andrzej Czyżewski
Gdańsk University of Technology
Multimedia Systems Department
andrzej.czyzewski@pg.edu.pl

Abstract

Federated learning is often framed as a practical trade-off in clinical NLP: safer data handling at the cost of lower predictive performance. We revisit this assumption in a benchmark-specific study of Polish medical text classification. A key issue is evaluation granularity: the test split contains 10,634 rows but only 670 unique normalized text hashes, with 18 inconsistent groups removed in strict grouped evaluation. We therefore compare centralized and federated training under both conventional instance-level scoring and a stricter hash-level protocol that controls duplicate inflation. In the strongest reported settings, federated training matches or slightly exceeds the centralized baseline, reaching instance-level Macro-F1 of 0.8826 ± 0.0177 versus 0.8689 ± 0.0124, and hash-level Macro-F1 of 0.8908 ± 0.0220 versus 0.8841 ± 0.0078. The claim is deliberately narrow: we do not argue that federated learning is generally superior to centralized training, nor do we claim formal privacy guarantees. Rather, we show that in this duplicate-heavy Polish medical text benchmark, conclusions about locality depend strongly on evaluation hygiene.

1 Introduction

Clinical natural language processing is increasingly treated as part of healthcare infrastructure rather than a purely academic benchmark exercise. Text classifiers can support document triage, coding assistance, cohort identification, quality control, and downstream analytics. In practice, however, the central bottleneck is often not model capacity but data movement. Raw medical text is difficult to centralize because of institutional boundaries, governance rules, privacy constraints, and the operational cost of maintaining trusted shared infrastructure. This motivates federated or locality-constrained learning, where model training remains compatible with institution-local data handling.

A common assumption follows from this setup: centralized training should usually outperform locality-constrained training because it sees all data in one place. Under that view, federated learning is mainly a governance concession rather than a genuinely competitive learning protocol. That intuition is plausible, especially in medical NLP, where class distributions can be uneven, labels noisy, and client-level data heterogeneous. Prior federated NLP and biomedical NLP studies also show that the centralized–federated gap is shaped by task type, partitioning, optimizer choice, and client heterogeneity rather than by a single universal locality penalty. The assumption can become especially misleading when evaluation is not aligned with the structure of the data being tested.

This paper studies one such source of distortion: duplication. In medical and adjacent text corpora, repeated text can arise from boilerplate formulations, templated fragments, dictated variants, recurring summary statements, source-specific formatting, or dataset construction procedures. These repetitions can cause the nominal size of a test set to overstate the amount of distinct textual evidence actually being evaluated. Under a standard instance-level protocol, models are rewarded on all rows independently, even when many rows correspond to the same underlying normalized text. As a result, comparisons between centralized and federated training may partly reflect duplicate structure rather than genuine differences in performance on distinct text units.

We therefore study federated versus centralized Polish medical text classification under a duplicate-aware evaluation protocol. The question is intentionally narrow: *does locality cost performance in this benchmark once duplicate inflation is controlled?* Our answer is not a general claim that federated learning is superior. Instead, we show that, in the strongest reported settings, the expected locality penalty is not observed. The best federated

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 498–508

models are competitive with the centralized baseline on conventional instance-level evaluation and slightly stronger on stricter hash-level evaluation, where duplicate texts are grouped and inconsistent duplicate groups are excluded.

The observed differences are modest, but the interpretation matters: once duplicate inflation is controlled, the usual locality penalty is not supported in the strongest settings. We therefore frame the paper as a methodological case study that compares centralized and federated learning across a structured sweep, reports both instance-level and hash-level evaluation, and inspects classwise behavior without claiming broad federated superiority.

Contributions.

- We present a benchmark-specific comparison of centralized and federated learning for Polish medical text classification under heavy duplication.
- We introduce a duplicate-aware hash-level evaluation protocol that groups repeated texts by normalized identity and excludes inconsistent duplicate groups.
- We show that, in this setting, the strongest federated configurations do not exhibit the expected locality penalty under either instance-level or hash-level Macro-F1.
- We argue that evaluation granularity materially affects how locality cost is interpreted in duplicate-heavy medical NLP benchmarks.

2 Related Work

Federated learning is commonly motivated in healthcare and biomedical NLP by privacy, governance, and data-access constraints that make raw-text centralization difficult or undesirable. In biomedical NLP, Peng et al. (2024) provide an in-depth evaluation of federated learning for information extraction and show that federated training can be competitive with centralized training in some settings, while remaining sensitive to federation design. At the benchmark level, Lin et al. (2022) introduce FedNLP and show that conclusions about federated NLP depend strongly on task type, partition strategy, and optimization method rather than on a simple centralized-versus-federated dichotomy.

A second recurring theme is heterogeneity. Realistic cross-silo setups are non-IID, and this heterogeneity can materially affect the gap between federated and centralized learning. In a real-world legal NLP benchmark, (Zhang et al., 2023b) show that natural client heterogeneity can preserve a noticeable performance gap relative to centralized training. Related evidence from multilingual depression detection likewise suggests that federated performance can approach centralized learning in some regimes while remaining sensitive to data distribution and training setup (Khalil et al., 2024). Taken together, these studies suggest that the cost of locality is not a universal constant but an empirical quantity shaped by federation realism.

Because client drift is central to cross-silo learning, the federated literature also emphasizes heterogeneity-aware optimization and aggregation. FedAvg (McMahan et al., 2017) remains the standard baseline in which clients perform local optimization followed by server-side parameter averaging, while FedProx (Li et al., 2020) adds a proximal term to reduce harmful local drift under heterogeneity. FedNLP compares standard baselines such as FedAvg and FedProx across NLP tasks and shows that optimizer choice matters, but not in a uniform way across settings (Lin et al., 2022). More generally, aggregation-focused work treats proximal or adaptive mechanisms as ways to stabilize training under statistical heterogeneity (Kairouz et al., 2021). For example, Palihawadana et al. (2022) study similarity-guided aggregation under non-IID conditions, while Zhang et al. (2023a) propose adaptive local aggregation for personalized federated learning. Complementarily, Sattler et al. (2023) show that the centralized–federated gap can also depend on what auxiliary information is available to the federation, not only on the choice of optimizer.

Our work is closest to this line of research, but differs in what it treats as the main methodological variable. Prior federated NLP work has focused primarily on task realism, partition realism, and optimizer behavior (Lin et al., 2022; Zhang et al., 2023b; Peng et al., 2024; Khalil et al., 2024). In contrast, we show that evaluation granularity itself can materially change the conclusion about locality in a benchmark with substantial duplication. Specifically, we compare centralized and federated training not only at the standard instance level, but also under a stricter hash-level protocol that groups duplicate texts and removes inconsistent duplicate groups. In a duplicate-heavy medical text setting, this evaluation choice is not cosmetic: it changes what counts as an independent test unit and therefore changes how the centralized–federated gap should be interpreted.

3 Task and Data

We study seven-way classification of Polish medical text. The focus is not on redefining the task itself but on measuring how training protocol and evaluation protocol interact under realistic duplication. The paper does not introduce a new dataset; instead, it re-examines an existing benchmark through the lens of duplicate-aware evaluation. We therefore focus on the dataset property that directly affects the central comparison: the gap between raw test rows and distinct normalized text identities.

The raw test split contains 10,634 rows. However, after grouping by normalized text identity, these rows collapse to only 670 unique text hashes. The average unique hash therefore corresponds to roughly 15.9 raw rows. We use *textual evidence* to refer to distinct normalized text identities rather than to the raw number of rows. This distinction is important because repeated rows can otherwise multiply the influence of the same underlying text on the reported metric.

This compression is not merely descriptive. In practice, repeated medical text can arise from templated documentation, repeated reporting phrases, dictated variants, source-specific formatting, or dataset construction procedures. If a test set is evaluated only at the raw instance level, repeated text may dominate the metric, and the resulting score can reflect the multiplicity of recurring phrases rather than robust performance on distinct content.

We therefore use two evaluation granularities:

- **Instance level**: every test row contributes independently.
- **Hash level**: texts are grouped by normalized text identity, and each unique group contributes at most one evaluation unit.

Duplicate definition. The hash-level protocol does not rely on semantic clustering or fuzzy near-duplicate detection. Duplicate groups are defined by deterministic normalized-text identity. Thus, a hash group corresponds to the same normalized text under the evaluation rule, not merely to examples that share a repeated phrase, boilerplate opening, or similar wording. This conservative definition avoids introducing an additional similarity threshold and makes the grouped evaluation reproducible.

Hash-level evaluation is stricter in two ways. First, it prevents repeated text from dominating the metric simply by appearing many times. Second, it makes label inconsistency explicit. Among the 670 unique test hashes, 18 are inconsistent, meaning that the same normalized text appears with conflicting labels. These groups are excluded from the strict hash-level evaluation, leaving 652 evaluated hashes. This exclusion is conservative: rather than forcing a majority vote over conflicting groups in the main experiment, we remove ambiguous units and measure performance on the remaining consistent text identities.

Table 1: Evaluation split statistics. Hash-level evaluation groups duplicate texts by normalized text identity and excludes inconsistent groups.

Statistic	Value
Test rows	10,634
Unique text hashes	670
Average rows per unique hash	15.9
Inconsistent hashes	18
Dropped inconsistent hashes	18
Evaluated hash groups	652
Number of classes	7

Table 1 summarizes the evaluation structure. The table is small, but it carries one of the most important points of the paper: benchmark difficulty and benchmark size are not well captured by raw test-row counts when duplication is heavy.

We use the term *locality-constrained learning* for the operational setting in which data remain institution-local during training. This phrasing is deliberate. The paper concerns decentralized data handling constraints and their impact on predictive quality. It is not a claim of formal differential privacy, secure aggregation, or cryptographic protection in the main experiments.

4 Methods

4.1 Modeling Setup

All main experiments use the same transformer classification family for seven-way Polish medical text classification. The central comparison in the paper is therefore not between architectures but between learning protocols evaluated under the same downstream metric. This design isolates the effect of locality and evaluation granularity rather than claiming an architecture-specific advance.

We consider three families of runs in the broader experiment space:

- **CENTRAL**: a pooled-data baseline trained conventionally.
- **FEDAVG**: a federated baseline with aggregation over local clients.
- **FEDPROX**: a proximal federated variant ex-

plored as an additional robustness-oriented alternative.

Federated optimization protocols. FedAvg is the standard federated baseline in which clients perform local optimization and the server periodically averages their model parameters (McMahan et al., 2017). FedProx extends this idea with a proximal term that discourages local client updates from drifting too far from the current global model, which can improve stability under heterogeneous client distributions (Li et al., 2020). In this paper, these protocols are used as learning-protocol baselines. They should not be interpreted as privacy mechanisms by themselves; the study evaluates predictive behavior under locality-constrained training, not formal privacy guarantees.

In the main results, the strongest configurations are CENTRAL and selected FEDAVG runs. FEDPROX remains part of the broader sweep and is informative about the search space, but it is not forced into the headline claim when it is not the strongest configuration.

4.2 Backbone Choice and Training Setup

The main experiments reported in this paper use `xlm-roberta-base` (Conneau et al., 2020) as the backbone encoder for seven-way Polish medical text classification. This choice is pragmatic rather than language-optimal. Our goal is not to identify the strongest Polish-specific encoder, but to compare centralized and federated learning under a stable and widely used pretrained transformer family that is easy to align across run types. This improves protocol consistency and comparability with broader federated NLP literature, where multilingual backbones are common.

At the same time, we do not treat this choice as neutral. A Polish-specialized model such as `allegro/herbert-base-cased` (Mroczkowski et al., 2021) is a legitimate alternative, and exploratory repository artifacts include such runs. However, the present paper does not claim a complete backbone-controlled comparison. The main claim is conditional on the reported XLM-R setup and should be interpreted as a study of locality under one fixed encoder family rather than as a definitive ranking of Polish-language pretrained models.

Inputs are tokenized with the corresponding pretrained tokenizer and truncated to a maximum sequence length of 96 tokens. Across the reported runs, the main optimization hyperparameters are stable: batch size 16, learning rate 2×10^{-5}, AdamW (Loshchilov and Hutter, 2019) optimization, and a linear warmup schedule. All headline results are reported over three random seeds.

Input-length protocol. The 96-token limit is treated as a fixed protocol choice applied identically to all centralized and federated runs. We do not claim that this value is the optimal truncation length for Polish medical text classification. Its role is to keep the backbone, preprocessing, and computational budget aligned across learning protocols, so that the main comparison isolates locality and evaluation granularity rather than sequence-length tuning.

The centralized baseline is trained for 3 epochs, and the checkpoint used for final test reporting is selected by best validation Macro-F1. For the three reported seeds, the selected checkpoints correspond to epochs 2, 3, and 3, respectively.

The federated runs are evaluated over 30 communication rounds with 1 local epoch per round. As in the centralized case, the final reported checkpoint for each seed is selected by best validation Macro-F1 (Opitz and Burst, 2019), but here at the communication-round level rather than the epoch level. For the headline FEDAVG configurations, the selected checkpoints occur at rounds 26, 25, and 30 for the $K = 5$ setting, and at rounds 30, 20, and 27 for the $K = 10$ setting.

4.3 Client Construction and Heterogeneity

Federated clients are created by Dirichlet partitioning (Osting and Reeb, 2017) of the training portion into K clients. Concretely, label-wise assignments are sampled from a Dirichlet distribution, with α used as the concentration parameter controlling client heterogeneity: lower α corresponds to more uneven and statistically heterogeneous client partitions, whereas larger α yields more homogeneous splits. The main paper reports the strongest locality-constrained configurations with $\alpha = 1.0$ and $K \in \{5, 10\}$. These are synthetic client partitions rather than natural institution-level splits, so they are used to study controlled locality effects rather than to claim deployment realism.

We retain μ in the configuration names to preserve direct alignment with the sweep artifacts and result tables. In the headline comparisons, the method identity follows the final summary tables, where the strongest configurations are labeled as FEDAVG.

4.4 Why Duplicate-Aware Evaluation Is Necessary

The justification for hash-level evaluation is methodological rather than cosmetic. If a test set contains many repeated normalized texts, then instance-level metrics effectively weight the same text identity by the number of times it appears. That may be acceptable for some operational use cases, but it is a weak default for judging whether a model has learned to classify *distinct medical text content*. In a paper about locality, this distinction becomes especially important because duplicated text identities can blur whether a system performs well on distinct evidence units or simply benefits from repeated lexical patterns.

Hash-level evaluation addresses this by collapsing duplicate rows into a single normalized text identity. It is stricter than raw instance-level evaluation and, in the present benchmark, meaningfully changes the interpretation of results. The centralized baseline is already strong on instance-level evaluation, but its score changes under hash-level grouping. Likewise, some federated configurations look better or worse depending on which granularity is used.

4.5 Selection of Headline Configurations

The repository contains a larger sweep than fits comfortably into the main narrative of the paper. To avoid turning the results section into a configuration dump, we use two principles for selecting headline rows:

1. report the centralized baseline;
2. report the strongest federated configurations under instance-level and hash-level Macro-F1.

This yields two useful headline federated configurations:

- **FEDAVG K5,** $\alpha = 1.0$**,** $\mu = 0.01$: strongest instance-level result;
- **FEDAVG K10,** $\alpha = 1.0$**,** $\mu = 0.01$: strongest hash-level result among the reported top settings.

The fact that the best instance-level and hash-level models are not identical is itself informative. It indicates that evaluation granularity affects which training setup appears strongest, reinforcing the paper's main methodological point.

5 Experimental Protocol

5.1 Primary Metric

We use Macro-F1 as the main metric. This choice reflects two considerations. First, the task has seven classes and class frequencies may differ, so raw accuracy alone would overweight larger or easier classes. Second, the main comparison in the paper is not about maximizing one benchmark number but about comparing protocols fairly across labels. Macro-F1 gives equal weight to each class and is therefore more appropriate than accuracy for protocol comparison under possible class imbalance. We consequently interpret the results as comparisons under a class-balanced metric, not as evidence that all classes are equally represented or equally easy.

5.2 Train, Validation, and Test Usage

All compared centralized and federated runs use the same underlying train, validation, and test division. The validation split is used for checkpoint selection, while the test split is reserved for final reporting. Federated clients are constructed from the training portion only; validation and test examples are not used in client construction. Thus, differences between CENTRAL and federated runs reflect the learning protocol and client partitioning rather than different evaluation data.

5.3 Seeds and Reporting

Headline results use three random seeds. We report means and standard deviations descriptively and avoid strong inferential claims from small differences.

5.4 Two Complementary Evaluation Views

All main comparisons are reported at both instance level and hash level:

- **Instance-level Macro-F1** reflects raw-row performance and is closer to a conventional benchmark view.
- **Hash-level Macro-F1** reflects distinct normalized-text performance and is more conservative in duplicate-heavy settings.

Neither view is meaningless, but they answer different operational questions. Instance-level evaluation is closer to a deployed row-stream setting, where repeated records may appear repeatedly and each row-level prediction can matter operationally. Hash-level evaluation is an audit view: it asks whether performance remains strong when repeated normalized texts no longer multiply their contribution to the metric. We therefore do not argue that instance-level evaluation is invalid. Rather, we argue that reporting only instance-level scores can obscure how much distinct textual evidence is actually being tested.

Table 2: Main comparison between the centralized model and the top federated configurations. Federated settings use $\alpha = 1.0$ and $\mu = 0.01$. Positive Δ indicates improvement over CENTRAL.

Model	Split	F1	Δ
CENTRAL	Inst.	0.8689±0.0124	–
	Hash	0.8841±0.0078	–
FedAvg K5	Inst.	**0.8826±0.0177**	0.0136
	Hash	0.8856±0.0214	0.0015
FedAvg K10	Inst.	0.8752±0.0198	0.0063
	Hash	**0.8908±0.0220**	0.0067

5.5 Interpreting Locality Cost

We use the term *locality cost* to denote the gap between centralized and federated performance. In the result tables, negative relative locality cost means that the federated configuration slightly exceeds the centralized baseline. For federated score F_{fed} and centralized score F_{cent}, the relative locality cost is reported as:

$$\text{LocCost}_{\text{rel}} = \frac{F_{\text{cent}} - F_{\text{fed}}}{F_{\text{cent}}}.$$

This convention is useful because it avoids presuming that locality must be costly. The sign of the quantity is allowed to be positive, zero, or negative.

6 Results

6.1 Main Comparison: Centralized vs Federated

Table 2 presents the central empirical result of the paper. The centralized baseline is strong, reaching instance-level Macro-F1 of 0.8689 ± 0.0124 and hash-level Macro-F1 of 0.8841 ± 0.0078. FEDAVG with $K = 5$, $\alpha = 1.0$, and $\mu = 0.01$ reaches the strongest instance-level result, 0.8826 ± 0.0177, which is an absolute difference of approximately $+0.0136$ relative to CENTRAL. FEDAVG with $K = 10$, $\alpha = 1.0$, and $\mu = 0.01$ reaches the strongest hash-level result, 0.8908 ± 0.0220, which is approximately $+0.0067$ above CENTRAL on the stricter grouped evaluation.

The headline interpretation is intentionally narrow: in this benchmark and under the reported settings, the expected locality penalty is not observed in the strongest configurations. This does not prove that federated training is universally better. It shows only that the simplest version of the centralized-superiority assumption is not supported here.

Table 3: Effect of evaluation granularity. Hash-level scores are computed after duplicate grouping and removal of inconsistent hash groups. Federated settings use $\alpha = 1.0$ and $\mu = 0.01$.

Model	Inst. F1	Hash F1	ΔF1
CENTRAL	0.8689±0.0124	0.8841±0.0078	0.0152
FedAvg K5	0.8826±0.0177	0.8856±0.0214	0.0030
FedAvg K10	0.8752±0.0198	0.8908±0.0220	0.0156

Table 4: Top federated configurations by instance-level Macro-F1. Negative relative locality cost means the federated model slightly outperforms CENTRAL.

Rank	K	α	μ	Inst. F1	Rel. cost
1	5	1.0	0.01	0.8826±0.0177	−0.0157±0.0075
2	10	1.0	0.01	0.8752±0.0198	−0.0072±0.0087
3	5	1.0	0.00	0.8752±0.0141	−0.0072±0.0043
4	10	1.0	0.00	0.8684±0.0182	0.0006±0.0077
5	5	0.3	0.01	0.8655±0.0102	0.0038±0.0225
6	5	0.3	0.00	0.8575±0.0067	0.0131±0.0096
7	20	1.0	0.01	0.8573±0.0059	0.0134±0.0098
8	20	0.3	0.00	0.8560±0.0154	0.0149±0.0102

6.2 Evaluation Granularity Changes the Story

The same results become more interpretable when viewed through the lens of evaluation granularity. Table 3 reports the shift from instance-level to hash-level Macro-F1. CENTRAL improves by $+0.0152$ when duplicates are grouped, and FEDAVG K10 improves by $+0.0156$. FEDAVG K5 changes much less, by about $+0.0030$.

This table does more than summarize numbers. It shows that model ranking depends on what one counts as an evaluation unit. If raw test rows are treated as independent evidence, FEDAVG K5 appears strongest. If unique normalized texts are treated as the more meaningful unit, FEDAVG K10 becomes more attractive. Duplicate-aware evaluation therefore does not merely rescale the scores; it can change which configuration one would choose.

6.3 Top Federated Configurations

To avoid over-interpreting a single favorable row, Table 4 reports the best federated configurations found in the broader sweep. The strongest entries all occur at high heterogeneity parameter $\alpha = 1.0$, and the top rows have negative relative locality cost. Under our sign convention, that means the federated model slightly exceeds the centralized baseline.

This broader view is useful for two reasons. First, it shows that the main claim is not based on one isolated row. Several federated configurations are very close to CENTRAL, and the strongest ones are slightly higher under the reported metric. Sec-

Table 5: Per-class F1 comparison between CENTRAL and FedAvg K10 ($\alpha = 1.0$, $\mu = 0.01$). Positive Δ favors the federated model.

Class	CENTRAL	FedAvg	Δ
0	0.9132±0.0094	0.9357±0.0312	**0.0225**
1	0.8037±0.0267	0.8270±0.0435	**0.0233**
2	0.8420±0.0143	0.8422±0.0239	0.0002
3	0.7948±0.0131	0.7977±0.0405	0.0030
4	0.9053±0.0190	0.9039±0.0268	-0.0014
5	0.9338±0.0153	0.9350±0.0074	0.0012
6	0.9959±0.0035	0.9939±0.0000	-0.0020

ond, it keeps the interpretation bounded: locality is not free under all settings. Some weaker or more difficult configurations remain worse. The correct conclusion is therefore not "federated always wins," but "the expected penalty of locality is configuration-sensitive and can disappear in the strongest settings."

6.4 Classwise Behavior

Aggregate Macro-F1 is useful, but it can hide pathological class behavior. A model can gain slightly overall while sacrificing one clinically meaningful label, and such a pattern would matter. We therefore compare CENTRAL with a strong federated configuration, FEDAVG K10 with $\alpha = 1.0$ and $\mu = 0.01$, at the per-class level.

Table 5 shows that the observed differences are not concentrated in one accidental category. Classes 0 and 1 show the strongest positive mean deltas, approximately $+0.0225$ and $+0.0233$. Classes 2, 3, and 5 are essentially neutral to slightly positive. Classes 4 and 6 show only small negative changes. The differences remain modest, but the pattern is not a single-class artifact.

The per-class table also helps explain why the result is interesting. If federated learning were merely trading off some classes against others in a brittle way, the aggregate comparison would be less convincing. Instead, the pattern is consistent with competitive performance across the label space under the reported protocol, while still requiring cautious interpretation because the number of seeds is small.

6.5 Uncertainty and Direction of Effect

One reasonable concern is whether the observed federated gains are just small random fluctuations. Table 6 therefore summarizes selected uncertainty statistics for the strongest configurations. Given the small number of seeds, we treat these summaries as descriptive rather than inferential. The important signal is not merely the point estimate but the direction of the seed-level deltas. For the strongest configuration, FEDAVG K5 with $\alpha = 1.0$ and $\mu = 0.01$, the federated model is non-inferior or better than CENTRAL in all three seeds under the stored direction summary.

Table 6: Selected uncertainty summaries for top federated settings. Negative Δ indicates a federated advantage over CENTRAL under the stored locality-cost convention.

K	α	μ	Abs.	Rel.	NP
Fed. advantage					
5	1.0	0.01	-0.0136 ± 0.0066	-0.0157 ± 0.0075	3/3
10	1.0	0.01	-0.0063 ± 0.0076	-0.0072 ± 0.0087	2/3
5	1.0	0.00	-0.0062 ± 0.0038	-0.0072 ± 0.0043	3/3
Near parity					
10	1.0	0.00	0.0005 ± 0.0066	0.0006 ± 0.0077	1/3

This is not a proof of universal superiority or statistical significance. It supports a narrower claim: in the best reported settings, federated performance is close to CENTRAL on average and directionally favorable across most or all seeds.

6.6 A Necessary Nuance: Locality Is Not Free Under Every Setting

A misleading interpretation would be that every federated configuration is strong. That is not true. In the wider sweep, harder heterogeneity settings can produce weaker and more variable outcomes. This matters because it frames the actual conclusion correctly.

The defensible version of the paper's claim is therefore not "locality never hurts." It is: *the cost of locality is not an inherent constant; it depends on how the federated problem is configured and on how evaluation handles duplicate structure.* This is the intended conclusion of the study.

7 Discussion

The paper has two messages: one empirical and one methodological.

The empirical message is modest. In this Polish medical text benchmark, the strongest federated configurations are competitive with the centralized baseline under the reported protocol. The centralized baseline is strong, and we do not minimize it. The result should therefore be read as evidence against an inevitable locality penalty in this benchmark, not as evidence that federated learning is generally superior.

The methodological message is more important. In duplicate-heavy medical text, evaluation design is not an afterthought. The difference between instance-level and hash-level reporting is not just a change in denominator; it changes what the model is being asked to prove. Instance-level evaluation rewards performance on the observed test rows, including repeated text. Hash-level evaluation rewards performance on distinct normalized content after collapsing duplicates and excluding inconsistent duplicate groups. Both views can be useful, but they answer different questions.

Practically, this suggests that duplicate-aware reporting should accompany, rather than replace, conventional row-level reporting in duplicate-heavy clinical NLP benchmarks. Row-level scores describe performance on the observed stream of records, including repeated templates and recurring formulations. Hash-level scores describe performance after repeated normalized texts no longer dominate the metric. Reporting both views makes it clearer whether an apparent gain reflects broad behavior over distinct content or stronger performance on frequently repeated text identities.

This framing also helps interpret small centralized–federated differences. In a narrow benchmark with three seeds, a small numerical advantage should not be treated as proof of algorithmic superiority. Its value is diagnostic: it shows whether the expected locality penalty remains visible after controlling the evaluation unit. Under this interpretation, the main finding is not that one protocol universally dominates the other, but that conclusions about locality are sensitive to benchmark structure and evaluation hygiene.

This distinction matters beyond the present benchmark. Many medical NLP datasets are assembled from heterogeneous sources and can inherit repeated phrasing, partial templates, or source-specific artifacts. If results are reported only at row level, then any conclusion about centralized versus federated learning may partly reflect the duplication profile of the corpus rather than a purely algorithmic difference.

A further reason the current result is useful is that the observed differences are not driven by a collapse in one or two classes. The per-class analysis shows modest positive differences in classes 0 and 1, near-neutral behavior in classes 2, 3, and 5, and only small negative differences in classes 4 and 6. This supports a cautious reading: the strongest federated configurations are competitive across the reported label space, but the evidence remains benchmark-specific and descriptive.

Finally, the broader sweep suggests a balanced interpretation. Federated learning is not automatically better. Some harder configurations remain weaker and more unstable. This is why the intended conclusion is not "federated wins," but rather that the presumed cost of locality is contingent and that duplicate-aware evaluation is necessary to measure it honestly.

Scope of the claim. The central claim is deliberately narrow. We do not conclude that federated learning is generally superior to centralized training in medical NLP. Nor do we conclude that the same behavior would necessarily hold for sequence labeling, question answering, generation, larger multilingual benchmarks, Polish-specialized encoders, or naturally institution-partitioned datasets. The evidence supports a more specific methodological point: in this duplicate-heavy Polish medical text classification benchmark, the expected centralized advantage is not observed once the comparison is reported under both row-level and duplicate-aware hash-level evaluation.

8 Conclusion

We revisited whether locality-constrained learning is necessarily worse than centralized training in a duplicate-heavy Polish medical text classification benchmark. Under both instance-level and hash-level evaluation, the strongest federated configurations match or slightly exceed the centralized baseline, but only within this narrow benchmark-specific setting.

The broader contribution is methodological. In duplicate-heavy medical text, evaluation granularity materially changes what benchmark numbers mean. Raw instance-level reporting can blur the difference between repeated phrasing and performance on distinct normalized text identities. Hash-level evaluation provides a stricter comparison by preventing repeated normalized texts from multiplying their contribution to the metric.

This does not establish federated superiority in general. It establishes that the cost of locality should be measured rather than presumed, and that duplicate-aware evaluation is a useful tool for doing so in benchmark settings of this kind.

9 Limitations

This study has several limitations.

First, the main empirical claim is benchmark-specific. We study one Polish medical text benchmark under one family of centralized and federated training runs. The paper therefore should not be read as evidence that federated learning generally outperforms centralized learning in clinical NLP. Rather, it shows that under heavy duplication and duplicate-aware evaluation, the expected locality penalty can disappear in this particular setting.

Second, the study is task-limited. We evaluate seven-way text classification, not named entity recognition, question answering, generation, or broader clinical NLP pipelines. These tasks may respond differently to federation because they depend on span boundaries, answer localization, entity frequency, longer contextual dependencies, or different annotation densities.

Third, the main text centers on `xlm-roberta-base` rather than a Polish-specialized encoder. This improves protocol consistency and comparability with broader federated NLP literature, but it limits conclusions about the strongest achievable Polish-language performance. The paper therefore does not claim that XLM-R is the best possible backbone for Polish medical text, only that it provides a consistent basis for the reported centralized–federated comparison.

Fourth, the client partitions are synthetic Dirichlet partitions rather than natural institution-level splits. They are useful for controlled comparison, but they are not a substitute for a multi-site clinical deployment study with naturally heterogeneous clients, local documentation practices, and site-specific label distributions.

Fifth, the reported results are based on three random seeds. This is sufficient for a minimal stability check, but not for strong inferential claims. We therefore treat the standard deviations as descriptive and interpret small gains cautiously.

Sixth, the paper is about locality-constrained learning, not formal privacy guarantees. We do not claim differential privacy, secure aggregation, or cryptographic privacy in the present study. Instead, we focus on performance under decentralized data handling constraints. In that sense, the present setup should be understood as complementary to stronger privacy-preserving mechanisms rather than as a replacement for them.

Seventh, hash-level evaluation depends on a specific normalization and grouping rule. Grouping by normalized text hash is well-motivated for this benchmark, but other corpora may require different definitions of textual identity. Exact normalized identity is conservative and reproducible, but it does not cover all possible forms of semantic near-duplication.

From an ethics perspective, text classifiers in medical settings should not be interpreted as standalone clinical decision systems. Their outputs are best treated as support signals embedded in broader human workflows. The main ethical contribution of this paper is methodological transparency: by making duplicate structure explicit, the work reduces the risk of overstating performance in a sensitive domain.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, and 1 others. 2021. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210.

Samar Samir Khalil, Noha S. Tawfik, and Marco Spruit. 2024. Federated learning for privacy-preserving depression detection with multilingual language models in social media posts. *Patterns*, 5(6):100990.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450.

Bill Yuchen Lin, Chaoyang He, Zihang Ze, H. Wang, Y. Hua, Christophe Dupuy, Rahul Gupta, Mahdi Soltanolkotabi, Xiang Ren, and Salman Avestimehr. 2022. FedNLP: Benchmarking federated learning methods for natural language processing tasks. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 157–175. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282. PMLR.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10. Association for Computational Linguistics.

Juri Opitz and Sebastian Burst. 2019. Macro F1 and macro F1. *arXiv preprint arXiv:1911.03347*.

Braxton Osting and Todd Harry Reeb. 2017. Consistency of dirichlet partitions. *arXiv preprint arXiv:1708.05472*.

Chamath Palihawadana, Nirmalie Wiratunga, Anjana Wijekoon, and Harsha Kalutarage. 2022. FedSim: Similarity guided model aggregation for federated learning. *Neurocomputing*, 483:432–445.

Le Peng, Gaoxiang Luo, Sicheng Zhou, Jiandong Chen, Ziyue Xu, Ju Sun, and Rui Zhang. 2024. An in-depth evaluation of federated learning on biomedical natural language processing for information extraction. *npj Digital Medicine*, 7:127.

Felix Sattler, Tim Korjakow, Roman Rischke, and Wojciech Samek. 2023. FedAux: Leveraging unlabeled auxiliary data in federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5531–5543.

Jianqing Zhang, Hua Yang, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. 2023a. FedALA: Adaptive local aggregation for personalized federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11237–11244.

Zhuo Zhang, Xiangjing Hu, Jingyuan Zhang, Yating Zhang, Hui Wang, Lizhen Qu, and Zenglin Xu. 2023b. FedLegal: The first real-world federated learning benchmark for legal NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 3492–3507. Association for Computational Linguistics.

A Appendix Overview

The appendix is intentionally compact. The main paper keeps a narrow narrative around centralized versus federated learning, instance-level versus hash-level evaluation, and classwise behavior. The appendix therefore includes only materials that directly support protocol transparency and reproducibility.

B Protocol Details

Table 7 summarizes the core experimental setup used by the main reported runs.

Table 7: Core protocol details for the main reported runs.

Field	Value
Main backbone	`xlm-roberta-base`
Exploratory alternative in repo	`allegro/herbert-base-cased`
Number of classes	7
Input truncation length	96
Batch size	16
Learning rate	2×10^{-5}
Optimizer	AdamW
Scheduler	Linear warmup
Central training length	3 epochs
Federated training length	30 rounds
Local training per round	1 local epoch
Seeds	3
Partitioning	Label-wise Dirichlet over K clients
Heterogeneity parameter	α (Dirichlet concentration)
Central model selection	Best validation Macro-F1 epoch
Federated model selection	Best validation Macro-F1 round

C Expanded Duplicate Audit

Table 8 reports the duplicate-aware test-set audit used by the hash-level evaluation protocol.

Table 8: Detailed duplicate-aware audit.

Statistic	Value
Test rows	10,634
Unique normalized hashes	670
Average rows per hash	15.9
Inconsistent hashes	18
Consistent hashes used	652

D Per-Seed Main Results

Table 9 reports the exact per-seed instance-level and hash-level Macro-F1 values for the run families used in the main text.

Table 9: Per-seed instance-level and hash-level Macro-F1 for the exact main-paper run families.

Model / Seed	Seed 1	Seed 2	Seed 3
CENTRAL inst. F1	0.8547	0.8769	0.8753
CENTRAL hash F1	0.8751	0.8890	0.8882
FEDAVG K5 inst. F1	0.8633	0.8980	0.8864
FEDAVG K5 hash F1	0.8611	0.9003	0.8954
FEDAVG K10 inst. F1	0.8524	0.8857	0.8876
FEDAVG K10 hash F1	0.8655	0.9016	0.9053

E Extended Uncertainty Summary

Table 10 reports configuration-level uncertainty summaries for selected federated settings.

Table 10: Configuration-level uncertainty summaries for selected federated settings.

K	α	μ	Abs.	Rel.	NP
5	1.0	0.01	-0.0136 ± 0.0066	-0.0157 ± 0.0075	3/3
10	1.0	0.01	-0.0063 ± 0.0076	-0.0072 ± 0.0087	2/3
5	1.0	0.00	-0.0062 ± 0.0038	-0.0072 ± 0.0043	3/3
10	1.0	0.00	0.0005 ± 0.0066	0.0006 ± 0.0077	1/3