MOSAIC: A Multilingual, Taxonomy-Agnostic, and Computationally Efficient Approach for Radiological Report Classification in Low-Resource Settings

Alice Schiavone[1,2], **Marco Fraccaro**[3], **Lea Marie Pehrson**[1,4,5], **Silvia Ingala**[2,4], **Rasmus Bonnevie**[3], **Michael Bachmann Nielsen**[1,4,5], **Vincent Beliveau**[7], **Melanie Ganz**[1,2], **Desmond Elliott**[1]

[1]Department of Computer Science, University of Copenhagen

[2]Neurobiology Research Unit, Copenhagen University Hospital

[3]Unumed ApS, [4]Department of Diagnostic Radiology, Copenhagen University Hospital

[5]Department of Clinical Medicine, University of Copenhagen

[5]Cerebriu A/S, [7]Institute for Human Genetics, Medical University of Innsbruck

Abstract

Radiology reports contain rich clinical information that can be used to train imaging models without relying on costly manual annotation. However, existing approaches face critical limitations: rule-based methods struggle with linguistic variability, supervised models require large annotated datasets, and recent LLM-based systems depend on closed-source or resource-intensive models that are unsuitable for clinical use. Moreover, current solutions are largely restricted to English and single-modality, single-taxonomy datasets. We introduce MOSAIC, a multilingual, taxonomy-agnostic, and computationally efficient approach for radiological report classification. Built on a compact open-access language model (MedGemma-4B), MOSAIC supports both zero-/few-shot prompting and lightweight fine-tuning, enabling deployment on consumer-grade GPUs. We evaluate MOSAIC across seven datasets in English, Spanish, French, and Danish, spanning multiple imaging modalities and label taxonomies. The model achieves a mean macro F1 score of 88 across five chest X-ray datasets, approaching or exceeding expert-level performance, while requiring only 24 GB of GPU memory. With data augmentation, as few as 80 annotated samples are sufficient to reach a weighted F1 score of 82 on Danish reports, enabling large-scale cohort classification with minimal human effort. Code and models are open-source, offering a practical alternative to large or proprietary LLMs in clinical settings.

Keywords: radiology reports classification, multilingual NLP, low-resource learning

1. Introduction

Deep learning methods have been extensively explored for AI-assisted medical imaging analysis. However, their effectiveness depends on large volumes of annotated data. Such annotations must be provided by expert radiologists, whose primary focus remains clinical care, limiting the availability of high-quality labeled datasets. A promising solution to the annotation bottleneck is to extract relevant information directly from radiology reports, which are routinely produced during imaging procedures to document abnormalities associated with clinical findings (Reichenpfader et al., 2024). This information can later be used to train medical imaging classifiers or perform retrospective clinical studies (Zhou et al., 2014). Classic methods to automate finding extraction from English radiology reports include rule-based methods and BERT-based classifiers, that can prove effective in providing annotations close to those of radiologists (Irvin et al., 2019; Smit et al., 2020). However, rule sets need to be hand-crafted and are still limited by syntactic variability, and deep learning methods need a large amount of annotations from expert clinicians (Yang et al., 2023). Adapting these methods to a new label taxonomy or language necessitates either retraining the model, partially or entirely, or developing new rule sets from scratch. Re-training models to adapt to a new language or taxonomy for the same underlying task is an inefficient use of computational resources and is not sustainable.

In contrast, recent advances in natural language processing have led to the emergence of large language models (LLMs), which follow user instructions through natural language prompting. These models enable zero-shot or few-shot classification (i.e., with no or very few labeled examples), eliminating the need for manual annotation or rule engineering, and offering greater adaptability across tasks, languages, and taxonomies (Gu et al., 2024; Dorfner et al., 2024). However, the use of LLMs in many clinical settings often depends on large and/or closed-source models that cannot be deployed locally, posing significant challenges for projects involving sensitive patient data due to privacy concerns and the need for high-end computational resources. In these scenarios, compact models that can operate on consumer-grade hardware are the most practical solution.

While several studies have explored this task, they are typically restricted to a single dataset, limited language coverage (often one or two languages), and limited to a single medical imaging

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 313–323

modality, such as X-rays or magnetic resonance imaging (MRI), each having distinct clinical contexts and reporting conventions (Reis et al., 2022; Nguyen et al., 2022; Mottin et al., 2023; Wollek et al., 2024; Olivato et al., 2024; Matsuo et al., 2024; Collado-Montañez et al., 2025; Mergen et al., 2025; Al Mohamad et al., 2025). Even when datasets share the same imaging modality, differences in research focus lead to variability in labeled findings. For example, a chest X-ray may be used to study either the heart or the lungs, resulting in highly diverse label sets across studies. While LLMs offer greater flexibility than traditional deep learning methods, they still face important limitations. Most models are primarily trained on English, as it dominates the available web-crawled data. As a result, despite their strong performance on many tasks, LLMs underperform compared to BERT-based approaches when applied to other languages, e.g. Danish MRI reports (Beliveau et al., 2024), or Japanese pancreatic cancer reports (Suzuki et al., 2024).

To identify a viable approach to radiological report classification, we investigate whether generative language models can effectively perform this task in multiple languages and taxonomies, particularly in low-resource settings with limited computational capacity, scarce annotated data, and the need for localized model deployment. We look at language and task competency, evaluating performance across diverse linguistic settings and label taxonomies, and analyzing trade-offs between model size, accuracy, and adaptability.

We propose MOSAIC, an efficient and flexible LLM-based method for radiological report classification with the following key properties:

- **Multilingual**: Trained and evaluated across multiple languages, including English, Spanish, French, and Danish.
- **Optimized for Small-scale**: Designed to be trained and tested on consumer-grade GPUs, enabling local model development and adaptation while preserving patient privacy by avoiding reliance on external services.
- **Adaptable**: Robust to variations in label taxonomies across diverse datasets and medical imaging modalities.
- **Annotation-efficient**: With data augmentation, the model can match full-dataset performance using as few as 80 annotated samples.
- **Open-source**: We release the code and models in two sizes (4B and 12B)[1], while most datasets are accessible through their original providers.

[1]Available upon paper acceptance

We invite researchers to evaluate our method on their own data. Additional datasets provided by the community will be incorporated in a future revision of this manuscript.

2. MOSAIC

MOSAIC is a language model specialized for prompt-based radiological report classification, available in either 4B and 12B versions. The final models are based on supervised fine tuning of the `MedGemma-4B` and `Gemma-12B` models using publicly available datasets. We describe our design decisions, including model selection, prompt design, fine-tuning strategies, and data augmentation to quantify the impact of each design choice.

2.1. Data

Few public radiological reports datasets are currently available, due to the risk of de-anonymization of patients or clinicians. Most reports are machine-annotated, which leads to noisy labels unsuitable for training language models. For high-quality results, we consider only datasets manually annotated or checked by radiologists. The key statistics of the datasets are presented in Table 1.

MIMIC-CXR (Johnson et al., 2019) is a collection of chest X-ray English radiological reports. It has annotations for 3 possible types of mentioned finding: positive mention ($+$), when a finding is described as present in the report; negative mention ($-$), as finding is described as absent in the report; and uncertain mention ($\sim$), when a definitive conclusion cannot be reached on a specific finding. All not mentioned findings in the report are assigned a "not mentioned" label. Only a small subset of the dataset was manually annotated and released.

PadChest-GR (Castro et al., 2024) is a manually-curated chest X-ray dataset. The original dataset was a collection of reports from a hospital in Spain. Later, the sentences describing abnormalities were extracted from text through LLM prompting, and translated to English. The extracted sentences were reviewed by a team of radiologists while inspecting the associated X-ray. We limit the label set to findings that occur at least 30 times.

CASIA-CXR (Metmer and Yang, 2024) is a French chest X-ray dataset. Each report was assigned one of five findings as positively mentioned. Labels for both *PadChest-GR* and *CASIA-CXR* can only be positively mentioned findings.

REFLACX (Bigolin Lanfredi et al., 2022) is based on *MIMIC-CXR*, from which unlabeled reports were extracted for manual annotation in two phases, with two different taxonomies: we refer to these as *REFLACX*^{I} and *REFLACX*^{II}. Each finding in these datasets was annotated as not mentioned, or

Dataset		Language	Modality	Number of Findings	Avg. Chars	Mention Classes	Train	Dev	Test
MIMIC-CXR	M	en	Chest X-Ray	14	760	$+, -, \sim$	535	50	100
PadChest-GR	P	es, en	Chest X-Ray	49	115	$+$	1951	100	879
CASIA-CXR	C	fr	Chest X-Ray	5	400	$+$	7677	100	3334
DanskCXR	D	da	Chest X-Ray	48	312	$+, -$	1600	125	750
Reflacx^{I}	R^{I}	en	Chest X-Ray	14	216	$+, \sim$	68	50	120
Reflacx^{II}	R^{II}	en	Chest X-Ray	15	201	$+, \sim$	1046	52	1098
DanskMRI	B	da	Brain MRI	3	1941	$+, -, \sim$	194	50	345

Table 1: Overview of the datasets used, including language, modality, number of findings, average characters per report, data split, and mention classes, namely positive $(+)$, negative $(-)$, and uncertain $(\sim)$ mentions of findings.

as mention with varying degrees of certainty, with a score from 1 to 5, following the definition by (Panicek and Hricak, 2016). As granular uncertainty detection is not the scope of this study, we map probable findings as positive mentions (score of 4 and 5), and otherwise uncertain.

DanskCXR (Anonymous, 2025) is a private chest X-ray dataset collected from hospitals and clinics in Denmark. Along with *MIMIC-CXR*, it is the only dataset that has annotations for negative mentions of findings. We limit our experiments to the 14 most frequent findings to exclude rare conditions.

We define as *DanskMRI* a dataset of MRI reports from Danish hospitals collected by Beliveau et al. (2024), who focused on identifying epilepsy-related findings in brain MRI radiology reports. MRI is a different medical imaging modality: the style and structure of MRI reports differ significantly from those of chest X-rays, presenting a distinct out-of-distribution scenario.

The dataset splits were generated using stratification (Sechidis et al., 2011) and implemented using the `iterative-stratification` Python library, trying to have a consistent number of validation samples across datasets.

2.2. Model selection and fine-tuning

We select a small language model to base MOSAIC on through an initial set of experiments comparing two model families at different sizes: Llama 3 (3B and 8B) (Grattafiori et al., 2024) and Gemma 3 (4B and 12B) (Team et al., 2025) model families. These models have similar computational requirements and inherent multilingual capabilities. Additionally, we include `Mmed-Llama-8B`, a multilingual medical foundation model based on `Llama-8B` (Qiu et al., 2024). Lastly, we test also three larger models of the same families (`Gemma-27B`, `MedGemma-27B` and `Llama-70B`). While these models are not feasible for local deployment on consumer-grade GPUs due to their high resource requirements, they serve to contextualize the performance of smaller alternatives. We select the instruction-tuned version of each model, given their general superiority for structured outputs (Zhang et al., 2025). Closed-source models are excluded, as they do not meet the crucial privacy requirements of this task.

Each model is fine-tuned on an NVIDIA RTX 4090 GPU (24GB) using its 4-bit quantized form along with Rank-Stabilized LoRA adapters, except for `Gemma-27B` and `Llama-70B` which are trained on a H100 (94GB). `Mmed-Llama-8B` does not have a 4-bit quantized model version. All the prompts and findings sets are written in English to leverage the models' stronger alignment to English instructions. The prompt structure is adapted based on the dataset's label taxonomy to ensure consistency with its definition of positive, uncertain, negative, and unmentioned findings. As an example, the prompt for *CASIA* is:

```
You are a helpful radiology assistant.
Given a radiology report, classify each
abnormality into a class. Output a valid
JSON with each abnormality as key, and the
class as value. The keys must be
['cardiomegaly', 'mass', 'pleural
effusion', 'pneumonia', 'pneumothorax'].
The values can be one of [-1, 1]. The
values have the following interpretation:
(1) the abnormality was mentioned, even
with uncertainty, in the report e.g. 'A
large pleural effusion', 'The cardiac
contours are stable.', 'The cardiac size
cannot be evaluated.'; (-1) the abnormality
was not mentioned in the report, or the
abnormality was negatively mentioned in the
report; e.g. 'No pneumothorax.'.
```

Fine-tuning is conducted using the Unsloth library (Han et al., 2023), while inference is performed with vLLM (Kwon et al., 2023). Before inference, the LoRA adapters are merged into the base models in 16-bit precision. Cross-entropy loss is used as the objective function. Fine-tuning and sampling parameters are selected based on empirical testing and a random hyperparameter search conducted using Weights & Biases (Biewald, 2020). Train-

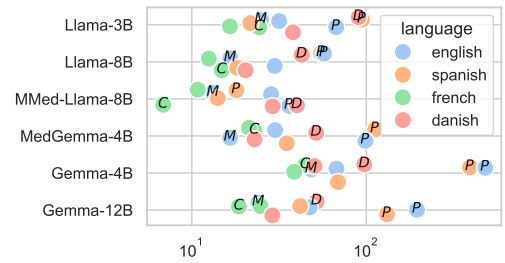

Figure 1: Perplexity scores on the SIB-200 and the chest X-ray reports, with the latter denoted by the dataset's initial.

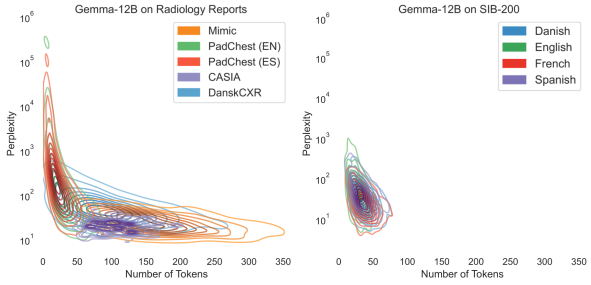

Figure 2: Distribution of perplexity scores as a function of sentence length on radiological reports and SIB-200, on model `Gemma-12B` .

ing configurations and hyperparameters are documented in the accompanying code repository.

2.3. Metrics

We use the F1 score to evaluate the extraction of findings for positive and negative mentions. The average of these scores yields the macro F1 score, reported as $(+)F1$ and $(-)F1$, respectively. To mitigate the impact of class imbalance, we report the weighted F1 score as $(w)F1$, which incorporates class support into the macro F1 calculation.

3. Results

To evaluate a model's ability to classify the radiology reports across multiple languages and taxonomies, we look at two key aspects: language competency and task competency.

3.1. Language competency

Language competency is the ability of a model to understand and generate text in different languages. It reflects how the model captures linguistic patterns and vocabulary, often measured by metrics like perplexity. Strong language competency enables better performance across multilingual tasks. SIB-200 is a large-scale, open-source benchmark for topic classification in 200 languages (Adelani et al., 2024). To assess language modeling capabilities, Figure 1 shows the distribution of perplexity scores on SIB-200 across the four focus languages in this paper and the chest X-ray reports. *MIMIC* and *CASIA*, along with their language counterparts in SIB-200, achieve lower perplexity scores than *DanskCXR* and *PadChest*. While Danish and Spanish datasets in SIB-200 generally yield higher perplexity, the radiology datasets exhibit particularly high scores. $PadChest_{en}$ also records higher perplexity than other English datasets. Among the models, `Mmed-Llama-8B` consistently attains the lowest perplexity, whereas the Gemma family yields the highest values, with `Gemma-4B` performing notably poorly on *PadChest*and `MedGemma-4B` showing a smaller spread. Figure 2 illustrates the relationship between text length and perplexity for `Gemma-12B`. SIB-200 shows a similar distribution, which does not seem affected by language. Source corpus strongly influences the perplexity score: in particular on radiological reports, *MIMIC* and $PadChest_{en}$, have very different distributions, even if both are written in English, confirming how language has a lower impact than dataset provenance in this task. Similar results were observed for other models but are omitted for brevity.

> ***Finding* 1.** Larger models yield lower perplexity, with domain-adapted variants providing additional gains. Dataset provenance also matters, with models performing better on data aligned with their training domain.

3.2. Task competency

Table 2 reports the performance of each model across five chest X-ray datasets under three experimental settings, evaluated using the $(+)F1$ score: zero-shot (ZS), where the model is directly prompted to classify the text according to the given taxonomy; three-shot (3S), using three examples randomly drawn from the corresponding training sets; and fine-tuned (FT) on *MIMIC*. In the 3-shot setting, three examples from the respective training set are provided. When fine-tuning, the models were instructed to output a valid JSON with each abnormality as key, and the mention type as the value, using a zero-shot prompt. The outputs are then parsed and validated.

A clear trend emerges across all models and datasets: performance improves with 3-shot prompting. Even if fine-tuning is performed only on one dataset, some models achieve an improvement over the other tested datasets. `Gemma-12B` consistently leads in performance, often achieving the highest F1 scores across multiple datasets and settings, particularly excelling in 3S and FT scenarios. `MedGemma-4B` also demonstrates strong

Dataset	MIMIC			PadChest$_{en}$			PadChest$_{es}$			CASIA			DanskCXR			*Average*		
Experiment	ZS	3S	FT	ZS	3S	FT	ZS	3S	FT	ZS	3S	FT	ZS	3S	FT	ZS	3S	FT
`Llama-3B`	46	53	77	60	66	67	39	53	53	55	75	79	53	53	57	50	60	66
`Llama-8B`	54	61	86	77	76	78	69	67	71	70	75	77	61	63	65	66	68	75
`Mmed-Llama-8B`	39	52	82	42	75	78	29	62	68	60	67	77	45	59	54	43	63	72
`Gemma-4B`	60	62	86	68	75	74	62	72	67	69	73	80	43	64	60	60	69	73
`MedGemma-4B`	55	59	88	61	77	74	53	72	71	69	82	82	62	65	58	60	69	75
`Gemma-12B`	65	70	84	79	79	80	76	76	76	76	76	81	69	75	69	73	75	78
`Gemma-27B` $*$	68	69	87	81	81	82	81	82	82	75	77	81	71	75	68	75	77	80
`MedGemma-27B` $*$	70	69	87	81	83	83	81	84	82	76	78	82	72	76	70	76	78	81
`Llama-70B` $*$	69	72	81	78	79	78	74	70	78	68	79	79	68	73	67	71	75	77

Table 2: Classification performance of language models on chest X-ray radiological free-text reports, measured in $(+)F1$ and ordered by model family and size. Models are tested under three settings: zero-shot (ZS); three-shot (3S) with three examples drawn from corresponding training sets; and the model fine-tuned on *MIMIC* (FT). We indicate with $(*)$ models fine-tuned on a 94GB H100, instead of a consumer-grade 24GB RTX 4090.

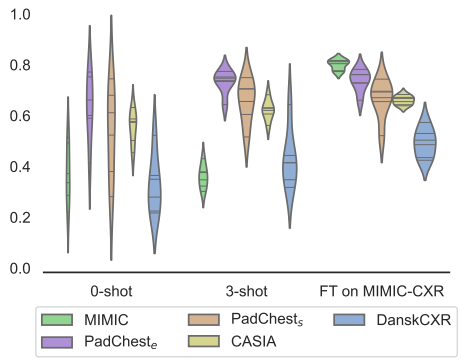

Figure 3: Distribution of $(+)F1$ scores from Table 2. Performance improves with increased supervision, but transferability varies by dataset.

performance, especially considering its smaller size, showing competitive F1 scores and even surpassing larger models in specific instances. `Llama-3B` and `Mmed-Llama-8B` generally show lower performance compared to Gemma variants, while `Llama-8B` performs generally well. Surprisingly, `Mmed-Llama-8B` performs poorly, especially compared to `Llama-8B`, particularly in zero-shot.

On the *MIMIC* dataset, CheXbert reports radiologist-level performance with a $(w)F1$ of 0.809, compared to 0.743 for rule-based methods and 0.798 for BERT-based classifiers (Smit et al., 2020). On a comparable subset, `MedGemma-4B` achieves $(w)F1$=88, suggesting that our approach can achieve expert-level performance. Large models (27B and 70B) have much higher scores on ZS and 3S, compared to the smaller ones. When fine-tuned, the performance is similar or equal on the target set, but these models better generalize on the test sets unseen during fine-tuning.

Figure 3 illustrates the distribution of the scores. While 0-shot learning shows a wide variability, the distributions become significantly tighter and higher for 3-shot learning. This trend is further pronounced when models are fine-tuned on *MIMIC*, suggesting a more robust and reliable performance. Interestingly, models consistently perform better on *PadChest$_{en}$* than on *PadChest$_{es}$*, despite both datasets containing the same reports but in different languages. This suggests that the task is easier in English than in Spanish. We conclude that larger models offer a good solution when no annotated data is available, but lack the accuracy of fine-tuning on the target data. Remarkably, in the fine-tuned setting smaller models match their performance, but show reduced generalization to out-of-distribution datasets. Table 3 shows runtime for training experiments on MIMIC, that we can compare to previous classification performance. Between the smaller models, `Gemma-12B` exhibits superior F1 scores but has higher memory and runtime demands. Additionally, training or inference may fail due to limited resources with large prompts (e.g. when the findings set or report is too long). `MedGemma-4B` stands out for its efficiency and competitive performance.

> ***Finding* 2.** In-context examples and supervised fine-tuning consistently improves performance compared to direct prompting. The best performing small-scale models are `MedGemma-4B` and `Gemma-12B`.

3.3. Multi-dataset fine-tuning

Table 4 reports $(+)F1$ scores using a series of increasingly comprehensive training setups, on the best-performing models in Section 3.2. Each successive row in the table reflects an incremental expansion of the training data, each indicated by the

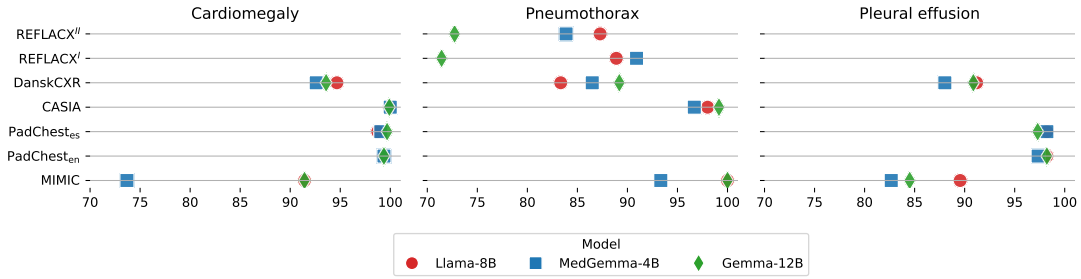

Figure 4: Performance as $(+)F1$ score of `MedGemma-4B`, `Llama-8B`, and `Gemma-12B` fine-tuned on MP$_{E+S}$C, with detailed results on three key Chest X-ray pathologies: Cardiomegaly, Pneumothorax, and Pleural Effusion. The data is presented across a range of chest X-ray datasets, illustrating model-specific and dataset-specific performance.

	Memory (GB)	Runtime (m)
`Llama-3B`	4.70	9.66
`Mosaic-4B`	8.69	13.98
`Gemma-4B`	9.16	13.88
`Llama-8B`	9.61	20.68
`Gemma-12B`	16.24	33.64
`Mmed-Llama-8B`	19.01	19.33
`Gemma-27B`	43.85	36.58
`MedGemma-27B`	41.48	34.65
`Llama-70B`	78.14	79.27

Table 3: Training runtime (in minutes) and peak GPU memory usage for models trained on *MIMIC*.

initial of the included dataset. The MP$_E$ setting includes only English-language data from *MIMIC* and *PadChest*$_{en}$, serving as the monolingual baseline. MP$_{E+S}$ adds Spanish samples, introducing cross-lingual variation. In MP$_{E+S}$C, French data from CASIA is incorporated, followed by the final setting MP$_{E+S}$CD, which adds Danish samples from DanskCXR, creating a multilingual, multi-institutional training configuration. We observe that the model trained under the MP$_E$ condition, using only English data, still achieves strong performance on *PadChest*$_{es}$, which contains the same reports as P$_E$ in Spanish. This result suggests that the model can generalize to the task itself without direct exposure to Spanish during fine-tuning, providing evidence of emerging *task competency*, even when the training data is exclusively monolingual, as long as the clinical structure remains consistent.

While the inclusion of additional languages and datasets does lead to gradual performance improvements, these gains are most pronounced in the final configuration MP$_{E+S}$CD, where all sources are present. The inclusion of *CASIA* in MP$_{E+S}$C yields near-to-perfect results on *CASIA*. This is not surprising, as this dataset is the largest (7.6k samples, compared to 0.5k of *MIMIC*) and with the simplest task, multi-class classification over 5 findings. Including *CASIA* and *DanskCXR* slightly degrades performance on *MIMIC* in smaller models, likely due to *MIMIC* becoming less proportionally represented in the combined training data. This effect is not observed in `Gemma-12B`. Figure 4 further illustrates this trend for three critical pathologies, observed across the relevant training sets and the English-language datasets *REFLACX*^{I} and *REFLACX*^{II}, under the MP$_{E+S}$CD setting. While `Gemma-12B` achieves the highest F1 scores, `MedGemma-4B` performs on par with or better than larger models in several scenarios, and on out-of-distribution *REFLACX* datasets.

> ***Finding* 3.** Adding more datasets improves performance, particularly in multilingual training. English-only models transfer to Spanish via shared clinical structure.

3.4. Taxonomy adaptation

Table 5 looks at the performance of `MOSAIC` trained on MP$_{E+S}$CD to assess task competency. The left columns (M and P$_E$) reflect the model's initial task competency on English-language findings. The red "X" marks in these columns indicate that these specific findings are *not present* in that dataset taxonomy. The right columns assess the model's task competency by measuring its ability to generalize to unseen English datasets R^{I}and R^{II}, both before and after fine-tuning on new data ($\Rightarrow$R). Although "Consolidation" is present in all datasets, fine-tuning on new data still yields substantial improvements. On findings that are present in only one of the training sets ("Nodule" and "Pneumothorax"), fine-tuning also significantly improves the models' ability to adapt their existing task competency to new distributions. However, for findings *not* included in the initial training set taxonomy, the fine-tuned model shows a small improvement ("Enlarger Hilum", of which there are 29 samples in the whole dataset) or, in the case of "Emphysema" (N=6), fine-tuning hurts performance.

	MIMIC			PadChest$_{en}$			PadChest$_{es}$			CASIA			DanskCXR			*Average*		
Datasets	MedGemma-4B	L1ama-8B	Gemma-12B	MedGemma-4B	L1ama-8B	Gemma-12B	MedGemma-4B	L1ama-8B	Gemma-12B	MedGemma-4B	L1ama-8B	Gemma-12B	MedGemma-4B	L1ama-8B	Gemma-12B	MedGemma-4B	L1ama-8B	Gemma-12B
M	87	86	84	74	78	80	71	71	76	82	77	81	58	65	69	74	75	78
MP$_E$	78	**89**	85	92	**95**	**95**	84	85	91	83	74	80	65	65	70	80	82	84
MP$_{E+S}$	80	82	84	93	94	**95**	92	**94**	**94**	83	76	79	61	61	69	82	81	84
MP$_{E+S}$C ⋆	76	85	84	92	**95**	94	89	**94**	**94**	**99**	**99**	**99**	66	63	69	84	87	88
MP$_{E+S}$CD	76	76	85	93	93	**95**	91	92	**94**	**99**	**99**	**99**	82	84	**86**	88	89	92

Table 4: $(+)F1$ scores with incrementally expanded multilingual training configurations on chest X-ray radiological reports. In bold, the best result over each dataset. MP$_E$ establishes a monolingual English baseline. Subsequent rows incrementally incorporate radiological reports in Spanish (P$_S$), French (C), and Danish (D), culminating in a multilingual, multi-institutional training setup. MedGemma-4B is 3 times smaller than Gemma-12B, but it still achieves a competitive performance. The symbol ⋆ indicates the training configuration used for the open-source MOSAIC-4B and MOSAIC-12B.

M	P$_E$		R$^{\text{I}}$	⇒R$^{\text{I}}$	R$^{\text{II}}$	⇒R$^{\text{II}}$
77	100	Consolidation	67	91	65	92
93	✗	Pneumothorax	90	100	84	93
✗	68	Nodule	47	60	✗	✗
✗	100	Hiatal Hernia	✗	✗	89	100
✗	✗	Emphysema	71	50	✗	✗
✗	✗	Enlarged Hilum	✗	✗	74	75

Table 5: Taxonomy adaptation in English of MOSAIC trained on MP$_{E+S}$CD, measured in $(+)F1$. Left columns show performance on present findings in training sets M and P$_E$; right columns show generalization to unseen datasets R$^{\text{I}}$ and R$^{\text{II}}$ before and after fine-tuning (⇒R).

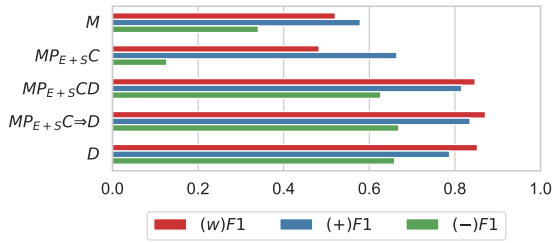

Figure 5: Evaluation of different pretraining and fine-tuning configurations of MOSAIC on the *DanskCXR*. Fine-tuning directly on D dataset outperforms zero-shot and task pretraining.

***Finding* 4.** Fine-tuning improves generalization to unseen datasets, especially for findings present in the training data, but offers limited gains for rare or missing findings.

3.5. Domain adaptation

DanskCXR is challenging out-of-distribution (OOD) training set for evaluating model adaptation beyond in-domain performance. This dataset is complex, as it is multi-class, multi-label, sourced from multiple institutions, and written in Danish. Notably, the current SOTA on this task is a BERT-based model that achieves a $(+)F1$ score of 88, across both positively and negatively mentioned findings. Figure 5 shows how different training strategies impact performance on *DanskCXR*. When pretrained on M or on MP$_{E+S}$C, the model struggles with negative finding detection: $(-)F1$=34, $(-)F1$=13, respectively. Incorporating *DanskCXR* into the pretraining phase yields no notable improvement over simply fine-tuning on *DanskCXR* alone. However, adapting a fine-tuned MOSAIC to *DanskCXR* (MP$_{E+S}$C⇒D) provides a modest performance boost, achieving $(+)F1$=84. We next ask: what is the minimal data requirement to match full-dataset performance on *DanskCXR*? As shown in Figure 6, we find that using as little as 30% of the dataset (480 examples) is sufficient to match the performance achieved with the full training set. To further examine data efficiency of small LLMs, Table 6 presents results when training is restricted to just 5% of the dataset (80 annotated examples).

***Finding* 5.** Domain adaptation performance is maximized by sequential adaptation to the target domain dataset. Direct pretraining on mixed datasets offers limited gains.

***Finding* 6.** Domain adaptation reaches its maximum performance with only 30% of the expert-annotated data (480/1600 examples).

	5%	10%	100%
D	75	80	85
D_e	73	74	85
D_{e+d}	79	80	86
$MP_{E+S}C \Rightarrow D$	75	82	87
$MP_{E+S}C \Rightarrow D_{e+d}$	82	84	86

Table 6: $(w)F1$ scores of MOSAIC trained on *DanskCXR* (D) and its English translation (D_e), with and without pretraining on $MP_{E+S}C$, on full data (100%) and two training dataset subsets (5% and 10%).

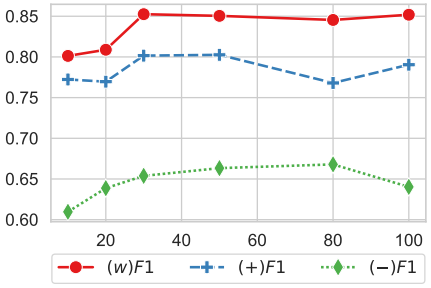

Figure 6: Data ablation study on *DanskCXR* shows that using just 480 examples (30% of the full dataset) is sufficient to reach peak performance.

3.6. Data augmentation

We also explore the benefit of data augmentation in domain adaptation. This experiment involves machine translating the original Danish dataset into English, obtained using `Gemma-27B`, to which we refer to as D_e. Models trained on this data were evaluated on the translated test set, which consistently yielded stronger results. Remarkably, initializing MOSAIC from $MP_{E+S}C$ and fine-tuning on only 5% of D achieves a $(w)F1$ score just 4 points below that of the full-data setup, highlighting the impact of pretraining and the surprisingly low data requirement for effective adaptation with data augmentation.

***Finding* 7.** Data augmentation through machine translation improves classification performance when there are limited resources in a desired language. This can reduce the effort required for expert data annotation.

3.7. Imaging modality adaptation

The *DanskMRI* dataset consists of Danish MRI reports annotated for three epilepsy-related brain abnormalities. Unlike chest X-ray datasets, these findings relate to neurological imaging, introducing both clinical and linguistic shifts. As shown in Figure 7, adaptive fine-tuning on external chest X-ray datasets ($MP_{E+S}C$) improves performance

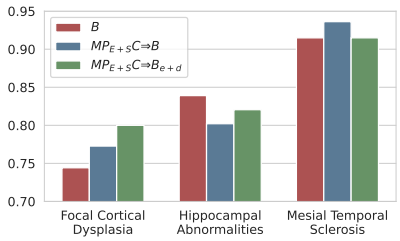

Figure 7: Performance of MOSAIC on the *DanskMRI* dataset, measured as $(+)F1$ across three epilepsy-related abnormalities from MRI reports.

over the base model for Focal Cortical Dysplasia and Mesial Temporal Sclerosis. This suggests that while cross-modality transfer can be effective, it may not generalize uniformly across all conditions. Adding English data augmentation improves consistency across all findings. In particular, it recovers performance on Hippocampal Abnormalities without sacrificing gains on the others. These results highlight the benefit of lightweight augmentation when adapting to new modalities, especially under language and data constraints, as only 194 examples are provided for fine-tuning.

***Finding* 8.** Cross-imaging modality transfer improves performance; English data augmentation enhances consistency.

4. Conclusion

We present MOSAIC, an approach for radiology report classification designed for local deployment and practical use. MOSAIC is a fine-tuned small language model trained on publicly available multilingual chest X-ray datasets (*MIMIC*, *PadChest*, *CASIA*). Unlike existing methods, it remains efficient while being flexible across languages and label taxonomies. We evaluate MOSAIC on English, Spanish, French, and Danish reports across two imaging modalities, finding robust performance in all settings. With as few as 80 annotated examples, it adapts effectively to new data distributions. As the first method to combine efficient multilingual adaptation with taxonomy-agnostic prompting, MOSAIC delivers state-of-the-art results across languages and taxonomies on standard hardware. By removing the need for large, expert-annotated datasets and language-specific models, it offers an accessible, scalable alternative for clinical NLP. MOSAIC is released in 4B and 12B parameter versions, both trained on public data, to support responsible and reproducible AI in healthcare.

Limitations

While MOSAIC shows strong performance across a range of datasets, it has several limitations. First, the model lacks interpretability: its predictions are based on internal transformer representations without transparent reasoning or token-level explanations. This opacity can be a barrier in clinical contexts, where trust and accountability are critical. Future work could explore integrating explainable components. Second, for some datasets there are no established baselines, making it challenging to benchmark relative performance. Third, our evaluation focuses primarily on positively mentioned findings. However, some clinical applications may require precise handling of negative and uncertain statements, which we do not evaluate on. Finally, although MOSAIC generalizes well across languages and imaging modalities, it struggles with rare or unseen conditions.

Acknowledgments

This research was funded by Innovation Fund Denmark grant number 0176-00013B and by Lundbeck Foundation grant (R279–2018–1145). Ethical approval for the dataset *DanskCXR* was obtained on 11 May 2022 from the Regional Council for Region Hovedstaden (R-22017450). Ethics approval for data collection for *DanskMRI* has been granted by the relevant ethics committees (CVK-2112460) and approval for data processing has been given by the Danish Data Protection Authorities.

5. Bibliographical References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2024. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects.

Fares Al Mohamad, Leonhard Donle, Felix Dorfner, Laura Romanescu, Kristin Drechsler, Mike P. Wattjes, Jawed Nawabi, Marcus R. Makowski, Hartmut Häntze, Lisa Adams, Lina Xu, Felix Busch, Aymen Meddeb, and Keno Kyrill Bressem. 2025. Open-source large language models can generate labels from radiology reports for training convolutional neural networks. *Academic Radiology*, 32(5):2402–2410.

Anonymous. 2025. Effective machine learning techniques for non-english radiology report classification: A danish case study. *AI*, 6(2).

Vincent Beliveau, Helene Kaas, Martin Prener, Claes N Ladefoged, Desmond Elliott, Gitte M Knudsen, Lars H Pinborg, and Melanie Ganz. 2024. Classification of radiological text in small and imbalanced datasets in a non-english language. *arXiv preprint arXiv:2409.20147*.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Ricardo Bigolin Lanfredi, Mingyuan Zhang, William F Auffermann, Jessica Chan, Phuong-Anh T Duong, Vivek Srikumar, Trafton Drew, Joyce D Schroeder, and Tolga Tasdizen. 2022. Reflacx, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. *Scientific data*, 9(1):350.

Daniel C Castro, Aurelia Bustos, Shruthi Bannur, Stephanie L Hyland, Kenza Bouzid, Maria Teodora Wetscherek, Maria Dolores Sánchez-Valverde, Lara Jaques-Pérez, Lourdes Pérez-Rodríguez, Kenji Takeda, et al. 2024. Padchest-gr: A bilingual chest x-ray dataset for grounded radiology report generation. *arXiv preprint arXiv:2411.05085*.

Jaime Collado-Montañez, María-Teresa Martín-Valdivia, and Eugenio Martínez-Cámara. 2025. Data augmentation based on large language models for radiological report classification. *Knowledge-Based Systems*, 308:112745.

Felix J Dorfner, Liv Jürgensen, Leonhard Donle, Fares Al Mohamad, Tobias R Bodenmann, Mason C Cleveland, Felix Busch, Lisa C Adams, James Sato, Thomas Schultz, et al. 2024. Is open-source there yet? a comparative study on commercial and open-source llms in their ability to label chest x-ray reports. *arXiv preprint arXiv:2402.12298*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jawook Gu, Han-Cheol Cho, Jiho Kim, Kihyun You, Eun Kyoung Hong, and Byungseok Roh. 2024. Chex-gpt: Harnessing large language models for enhanced chest x-ray report labeling. *arXiv preprint arXiv:2401.11505*.

Daniel Han, Michael Han, and Unsloth team. 2023. Unsloth. http://github.com/unslothai/unsloth. Software.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik

Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Alistair Johnson, Matt Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg-chest radiographs with structured labels. *PhysioNet*, 101:215–220.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Hidetoshi Matsuo, Mizuho Nishio, Takaaki Matsunaga, Koji Fujimoto, and Takamichi Murakami. 2024. Exploring multilingual large language models for enhanced tnm classification of radiology report in lung cancer staging. *Cancers*, 16(21):3621.

Markus Mergen, Daniel Spitzl, Conrad Ketzer, Maximilian Strenzke, Alexander W. Marka, Marcus R. Makowski, Keno K. Bressem, Lisa C. Adams, and Florian T. Gassert. 2025. Leveraging large language models for accurate ao fracture classification from ct text reports. *Journal of Imaging Informatics in Medicine*.

Hichem Metmer and Xiaoshan Yang. 2024. An open chest x-ray dataset with benchmarks for automatic radiology report generation in french. *Neurocomputing*, 609:128478.

Luc Mottin, Jean-Philippe Goldman, Christoph Jäggli, Rita Achermann, Julien Gobeill, Julien Knafou, Julien Ehrsam, Alexandre Wicky, Camille L Gérard, Tanja Schwenk, et al. 2023. Multilingual recist classification of radiology reports using supervised learning. *Frontiers in digital health*, 5:1195017.

Thao Nguyen, Tam M Vo, Thang V Nguyen, Hieu H Pham, and Ha Q Nguyen. 2022. Learning to diagnose common thorax diseases on chest radiographs from radiology reports in vietnamese. *Plos one*, 17(10):e0276545.

Matteo Olivato, Luca Putelli, Nicola Arici, Alfonso Emilio Gerevini, Alberto Lavelli, and Ivan Serina. 2024. Language models for hierarchical classification of radiology reports with attention mechanisms, bert, and gpt-4. *IEEE Access*, 12:69710–69727.

David M Panicek and Hedvig Hricak. 2016. How sure are you, doctor? a standardized lexicon to describe the radiologist's level of certainty. *American Journal of Roentgenology*, 207(1):2–3.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384.

Daniel Reichenpfader, Henning Müller, and Kerstin Denecke. 2024. A scoping review of large language model based approaches for information extraction from radiology reports. *npj Digital Medicine*, 7(1).

Eduardo P Reis, Joselisa PQ De Paiva, Maria CB Da Silva, Guilherme AS Ribeiro, Victor F Paiva, Lucas Bulgarelli, Henrique MH Lee, Paulo V Santos, Vanessa M Brito, Lucas TW Amaral, et al. 2022. Brax, brazilian labeled chest x-ray dataset. *Scientific Data*, 9(1):487.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*, pages 145–158. Springer.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. 2020. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. *CoRR*, abs/2004.09167.

Kazufumi Suzuki, Hiroki Yamada, Hiroshi Yamazaki, Goro Honda, and Shuji Sakai. 2024. Preliminary assessment of tnm classification performance for pancreatic cancer in japanese radiology reports using gpt-4. *Japanese Journal of Radiology*, 43(1):51–55.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Alessandro Wollek, Sardi Hyska, Thomas Sedlmeyr, Philip Haitzer, Johannes Rueckel, Bastian O. Sabel, Michael Ingrisch, and Tobias Lasser. 2024. German chexpert chest x-ray radiology report labeler. *RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, 196(09):956–965.

Eric Yang, Matthew D Li, Shruti Raghavan, Francis Deng, Min Lang, Marc D Succi, Ambrose J Huang, and Jayashree Kalpathy-Cramer. 2023. Transformer versus traditional natural language processing: how much data is enough for automated radiology report classification? *The British Journal of Radiology*, 96(1149):20220769.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. 2025. Gpt4roi: Instruction tuning large language model on region-of-interest. In *European Conference on Computer Vision*, pages 52–70. Springer.

Yirong Zhou, Paul K Amundson, Fangsheng Yu, Matthew M Kessler, Tammie L S Benzinger, and Franz J Wippold. 2014. Automated classification of radiology reports to facilitate retrospective study in radiology. *Journal of Digital Imaging*, 27(6):730–736.