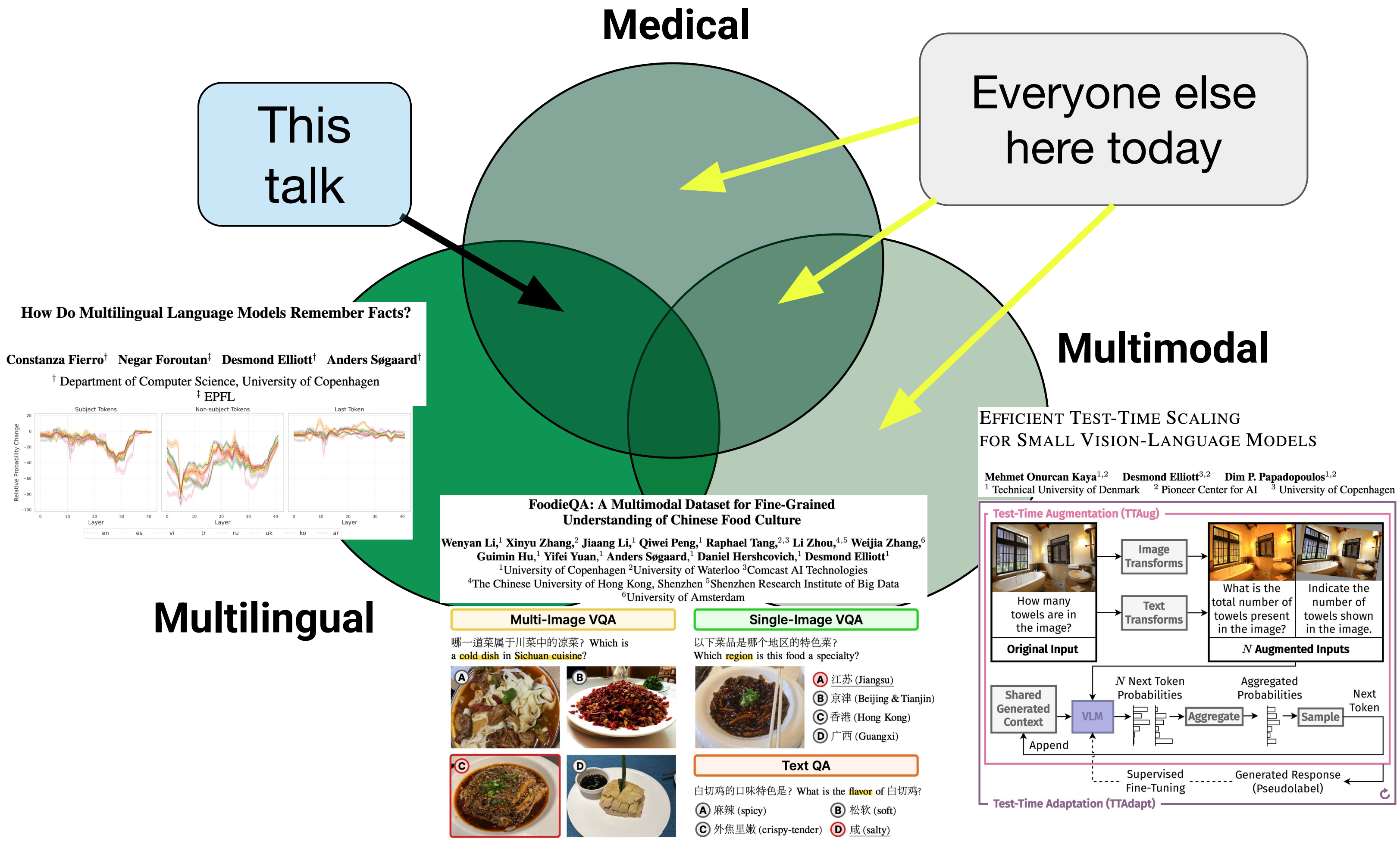
Multilingual Radiology Report Classification

Desmond Elliott

Language and Multimodal Processing Group
Department of Computer Science
University of Copenhagen

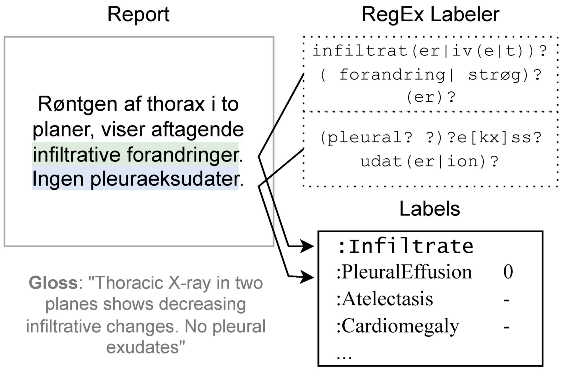

Medical
This talk
Everyone else here today
Multimodal
Multilingual
How Do Multilingual Language Models Remember Facts?
Constanza Fierro† Negar Foroutan‡ Desmond Elliott† Anders Søgaard†
† Department of Computer Science, University of Copenhagen
‡ EPFL
Subject Tokens
Non-subject Tokens
Last Token
Relative Probability Change
Layer
en es vi tr ru uk ko ar
FoodieQA: A Multimodal Dataset for Fine-Grained Understanding of Chinese Food Culture
Wenyan Li,[1] Xinyu Zhang,[2] Jiaang Li,[1] Qiwei Peng,[1] Raphael Tang,[2,3] Li Zhou,[4,5] Weijia Zhang,[6] Guimin Hu,[1] Yifei Yuan,[1] Anders Søgaard,[1] Daniel Hershcovich,[1] Desmond Elliott[1]
[1]University of Copenhagen [2]University of Waterloo [3]Comcast AI Technologies
[4]The Chinese University of Hong Kong, Shenzhen [5]Shenzhen Research Institute of Big Data
[6]University of Amsterdam
Multi-Image VQA
哪一道菜属于川菜中的凉菜？Which is a cold dish in Sichuan cuisine?
Single-Image VQA
以下菜品是哪个地区的特色菜？Which region is this food a specialty?
A 江苏 (Jiangsu)
B 京津 (Beijing & Tianjin)
C 香港 (Hong Kong)
D 广西 (Guangxi)
Text QA
白切鸡的口味特色是？What is the flavor of 白切鸡?
A 麻辣 (spicy)
B 松软 (soft)
C 外焦里嫩 (crispy-tender)
D 咸 (salty)
EFFICIENT TEST-TIME SCALING FOR SMALL VISION-LANGUAGE MODELS
Mehmet Onurcan Kaya[1,2] Desmond Elliott[3,2] Dim P. Papadopoulos[1,2]
[1] Technical University of Denmark [2] Pioneer Center for AI [3] University of Copenhagen
Test-Time Augmentation (TTAug)
How many towels are in the image?
Original Input
Image Transforms
Text Transforms
What is the total number of towels present in the image?
Indicate the number of towels shown in the image.
N Augmented Inputs
Shared Generated Context
Append
VLM
N Next Token Probabilities
Aggregate
Aggregated Probabilities
Sample
Next Token
Supervised Fine-Tuning
Generated Response (Pseudolabel)
Test-Time Adaptation (TTAdapt)

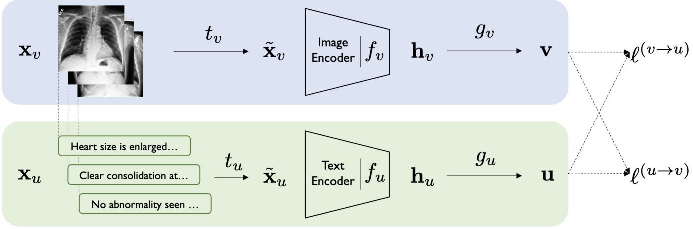
Radiology Report Classification

- Backbone of training imaging classification systems
  - `regex` is everywhere
  - SSL emerging as an alternative
    - This is compute-intensive compared to LLM knowledge

- Not much publicly shared data
  - MIMIC-CXR, CheXpert, etc.

- Disjoint findings labels
  - MIMIC-CXR: 15 findings
  - PadChest: 49 findings





How can we combine publicly available radiology report resources into a single classification model?

Alice

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

Alice Schiavone[1,2], Marco Fraccaro[3], Lea Marie Pehrson[1,4,5], Silvia Ingala[4,6], Rasmus Bonnevie[3]
Michael Bachmann Nielsen[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
[6]Cerebriu A/S, [7]Institute for Human Genetics, Medical University of Innsbruck

Desiderata

- **Fully open source:** keep your medical data on-site
- **Accessible:** run and train on inexpensive general-purpose GPUs
  - training and inference on an 24GB RTX 3090
- **Multilingual:** works for any EU26 major language
  - only evaluated in 4 languages due to data availability
- **Flexible:** adapts to different findings labels with minimal intervention

The LLM money pit

The rich man experience:

- LLM can solve 6 additional **high-school** competition math problems (AIME) for 4.2M USD
    - Reaching 65% on the test (below the acceptance cutoff) using only 16 tries…

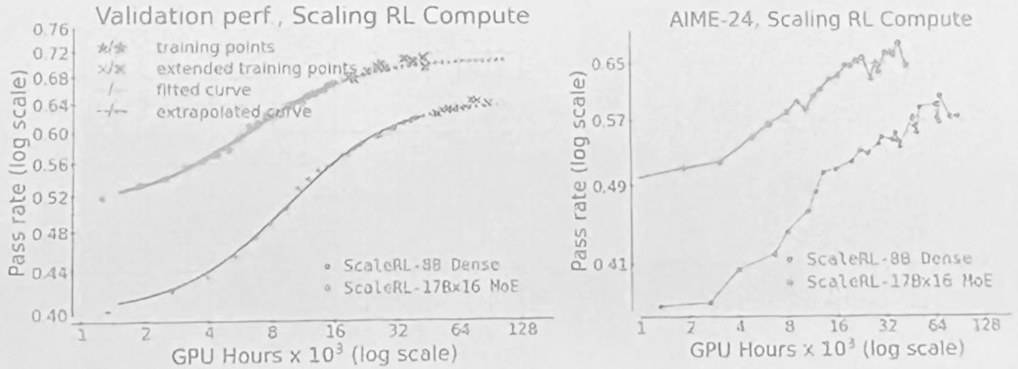
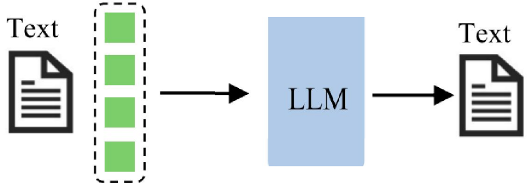


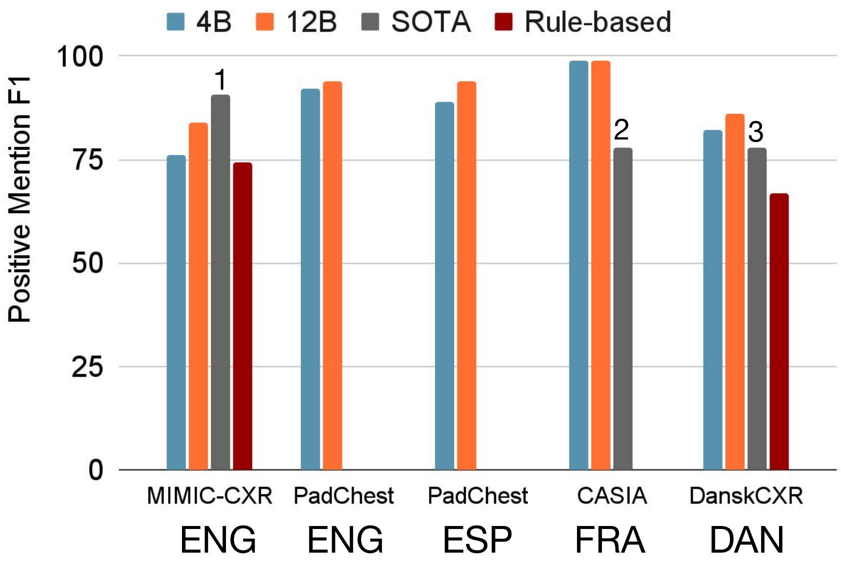
Figure 1 Predicatably Scaling RL compute to 100,000 GPU Hours (a) We run **ScaleRL** for 100k GPU hours on an 8B dense

Alexia Jolicoeur-Martineau, Mila, October 2025

MOSAIC-4B and 12B

- Finetuned on 10K reports in English, Spanish, and French
  - QLoRA optimization on the Q,K,V, FF, and Output layers
  - Need maximum of 16.2G VRAM and 33 minutes for SFT
- Prompt-based inference that can predict up to 68 findings





1. CheX-GPT, 2. CASIA-CLS, 3. DanskBERT

Lucas Dixon, Google DeepMind

How ~~I~~ think about LLMs...

An interpreter (that can translate between languages, concepts, and styles)

An improv comedian

A fuzzy database of the web

Prompt-based Inference Example

Require JSON-structured responses

```
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
{findings}. The values can be one of
{classes}. The values have the following
interpretation:
```

Define style of positive/uncertain findings

```
(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.';
```

Negative mentions

```
(2) the abnormality was negatively
mentioned in the report; e.g. 'No
pneumothorax.'
```

Datasets

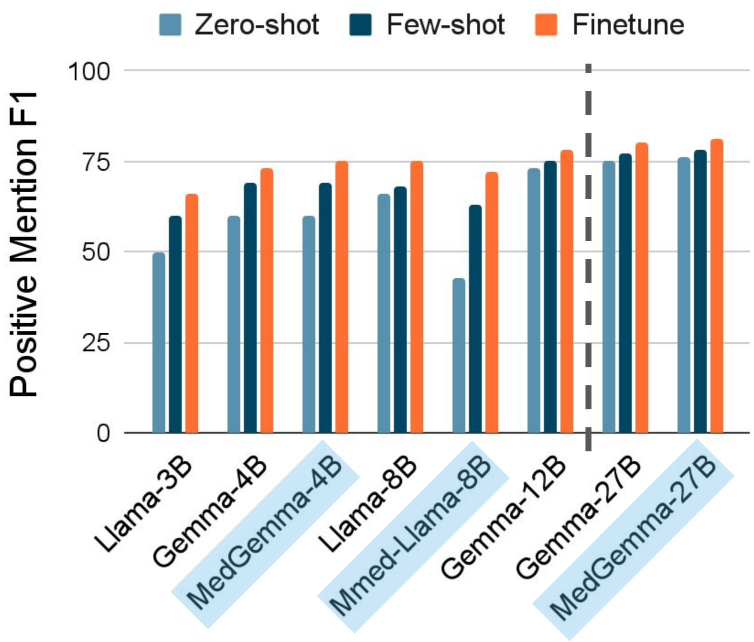
Dataset	Language	Modality	Number of Findings	Avg. Chars	Mention Classes	Train	Dev	Test
MIMIC-CXR	en	Chest X-Ray	14	760	$+, -, \sim$	535	50	100
PadChest-GR	es, en	Chest X-Ray	49	115	$+$	1951	100	879
CASIA-CXR	fr	Chest X-Ray	5	400	$+$	7677	100	3334
DanskCXR	da	Chest X-Ray	48	312	$+, -$	1600	125	750
DanskMRI	da	Brain MRI	3	1941	$+, -, \sim$	194	50	345

- Focus on publicly available datasets
    - 194-7600 training examples
    - 115–1941 characters
    - 3–49 findings across variable number of mention classes
- DanskMRI evaluates performance on different imaging modality

Which Backbone LLM?

- Setups:
  - Zero-shot prompting
  - Few-shot prompting
  - Dataset-specific fine-tuning

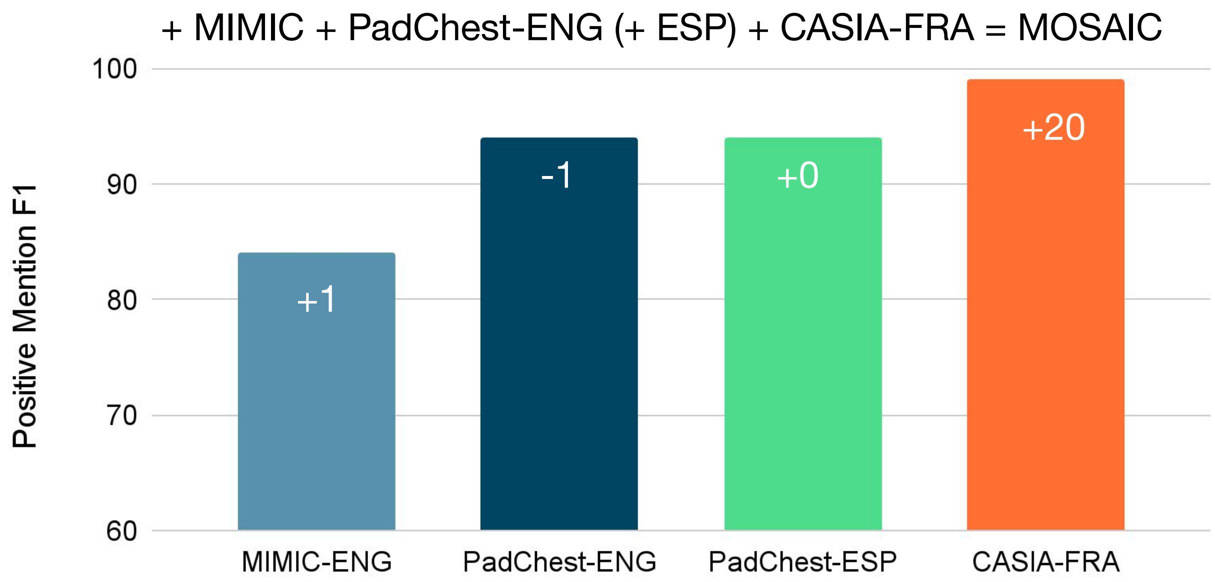
- Gemma and LLaMA LLMs
  - 3B–27B variants
  - General and medical domain



Finding 1: No substantial difference between general / medical domain models

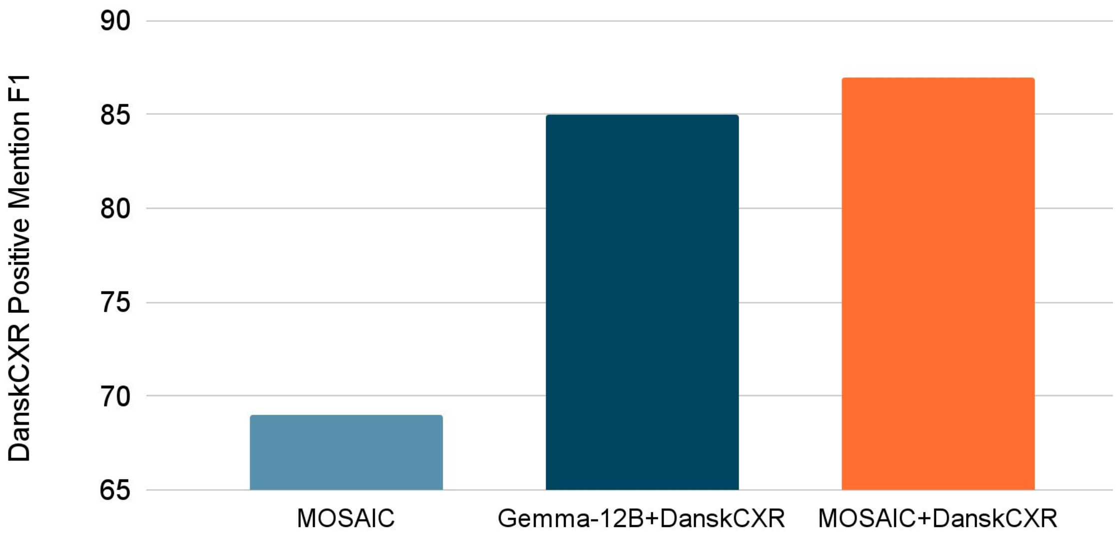
Finetuning on Public Datasets

- How does performance improve as we train on different label sets?



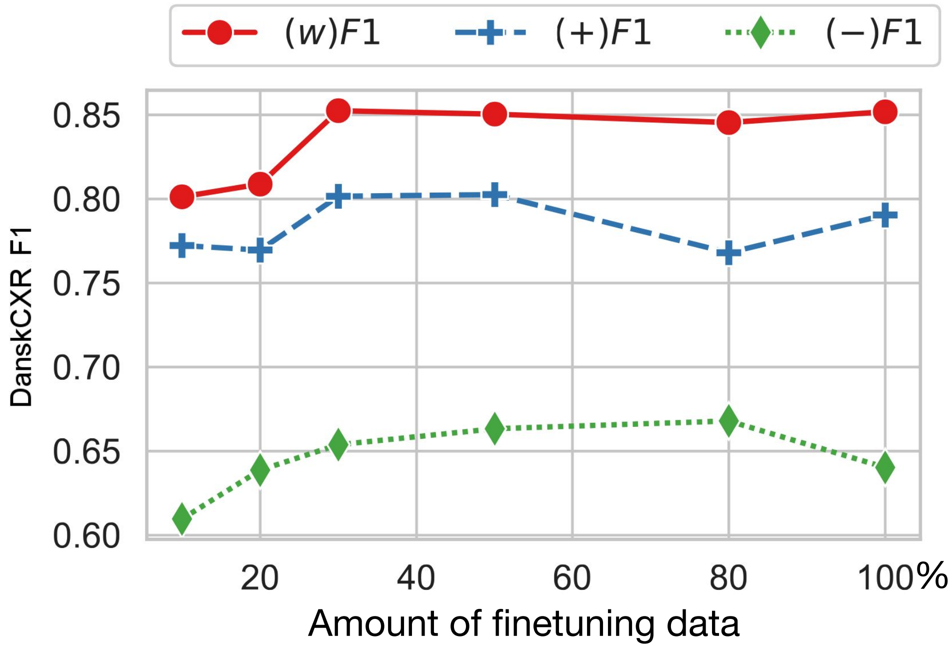
Finding 2: Improvements are additive and do not seem to interfere

New Dataset Adaptation



Finding 3: MOSAIC is a better starting point for new data

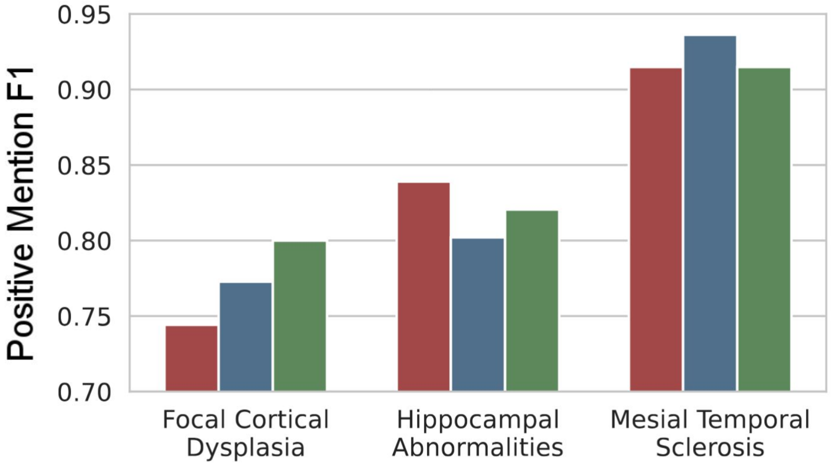
How Much Data Do You Need?



Finding 4: You don't need much data if you start from MOSAIC

Different Imaging Modality

- Adapt MOSAIC to predicting three findings in Epilepsy MRI reports
  - Gemma-12B SFT on 196 reports
  - MOSAIC-12B
    - + ENG translation of 196 reports



Finding 5: MOSAIC can be repurposed to a new modality

Open Directions

- **Multimodal inputs** could improve performance but how to handle reports from different imaging modalities
- **Simple text-only augmentation** could substantially improve performance [Aepli and Sennrich, 2022; Kaya et al. 2025]
- **Multi-agent LLMs** could better handle different mention classes
- **Broken tokenizers** could be fixed to further improve performance
    - See, e.g. TokenDist [Dobler et al. 2025]
- **Synthetic data generation** using self-consistency [Wang et al. 2023]

Conclusions

- Multilingual LLMs are radiology report classifiers
  - Handle different label sets
  - Handle reports from different imaging modalities
- Multilingual multi dataset SFT can reduce the total amount of data that needs expert annotation
  - Focus the time of our clinical colleagues on labelling lower-frequency findings or difficult examples
- MOSAIC is open source
  - Please tell us if it works for your data and language

References

- A. Schiavone M. Fraccaro, L. M. Pehrson, S. Ingala, R. Bonnevie, M. B. Nielsen, V. Beliveau, M. Ganz, and D. Elliott. 2025. MOSAIC: A Multilingual, Taxonomy-Agnostic, and Computationally Efficient Approach for Radiological Report Classification
- K. Dobler, D. Elliott, and G. de Melo. 2025. Token Distillation: Attention-aware Input Embeddings For New Tokens.
- M. O. Kaya, D. Elliott, and D. P. Papadopoulos. 2025. Efficient Test-Time Scaling for Small Vision-Language Models.
- Aepli and Sennrich. ACL 2022. Improving Zero-Shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-Level Noise.
- Wang et al. ICLR 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models.