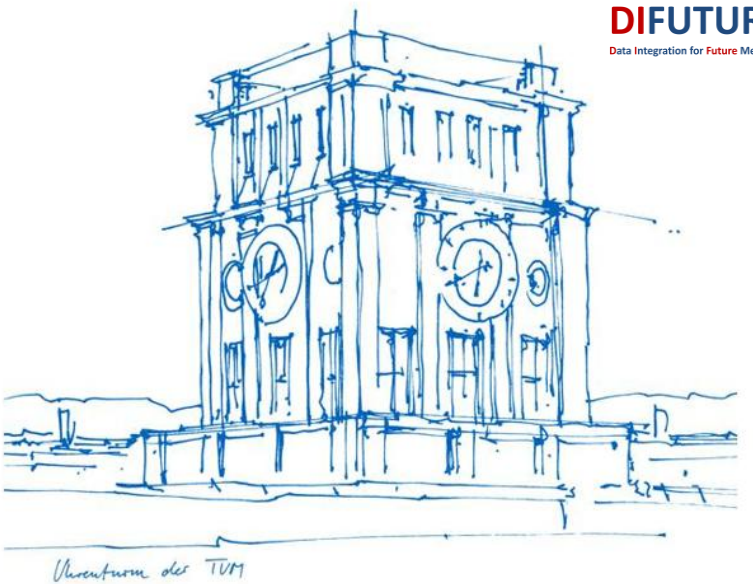
German Medical NER with BERT and LLMs: The Impact of Training Data Size

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
DIFUTURE
Data Integration for Future Medicine

Suteera Seeha

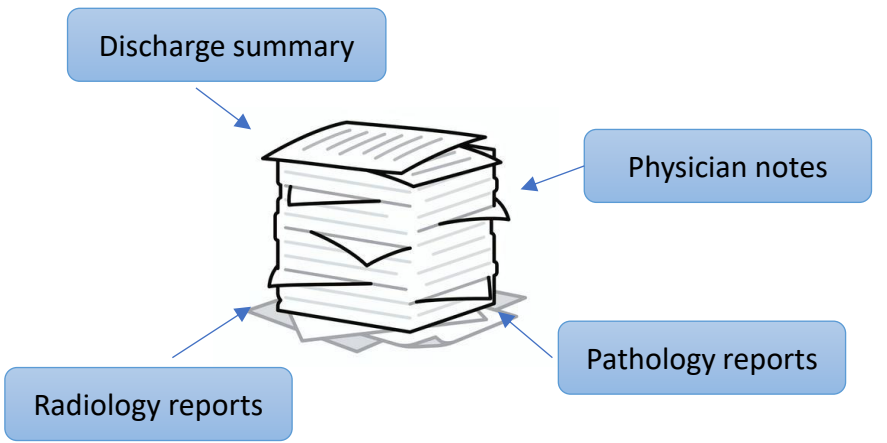
Sihan Wu, Justin Hofenbitzer, Claudio Benzoni, Peter Pallaoro, Raphael Scheible, Martin Boeker, Luise Modersohn

MIE, May 2025

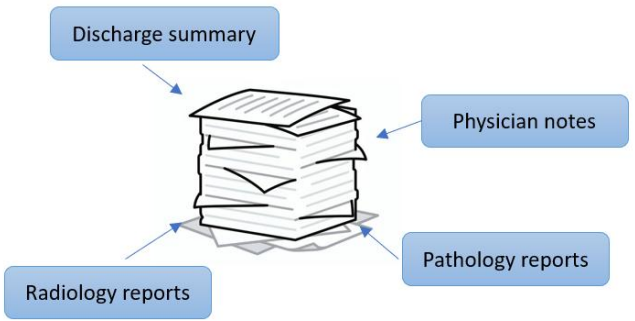


Institute of AI and Informatics in Medicine (AIIM), TUM University Hospital,
Technical University of Munich, Munich, Germany

Unstructured data is often underutilized



Unstructured data contain hidden insight



- Integrating relevant information to EHR or / data integration center
- Clinical decision support systems
- Medical coding (e.g., ICD-10 coding)

Named Entity Recognition (NER)

symptom symptom symptom

The patient reported experiencing itchy skin, runny nose, and watery eyes for the past three days.

medication

She was prescribed Cetirizine, 10 mg, to be taken once daily.

diagnosis

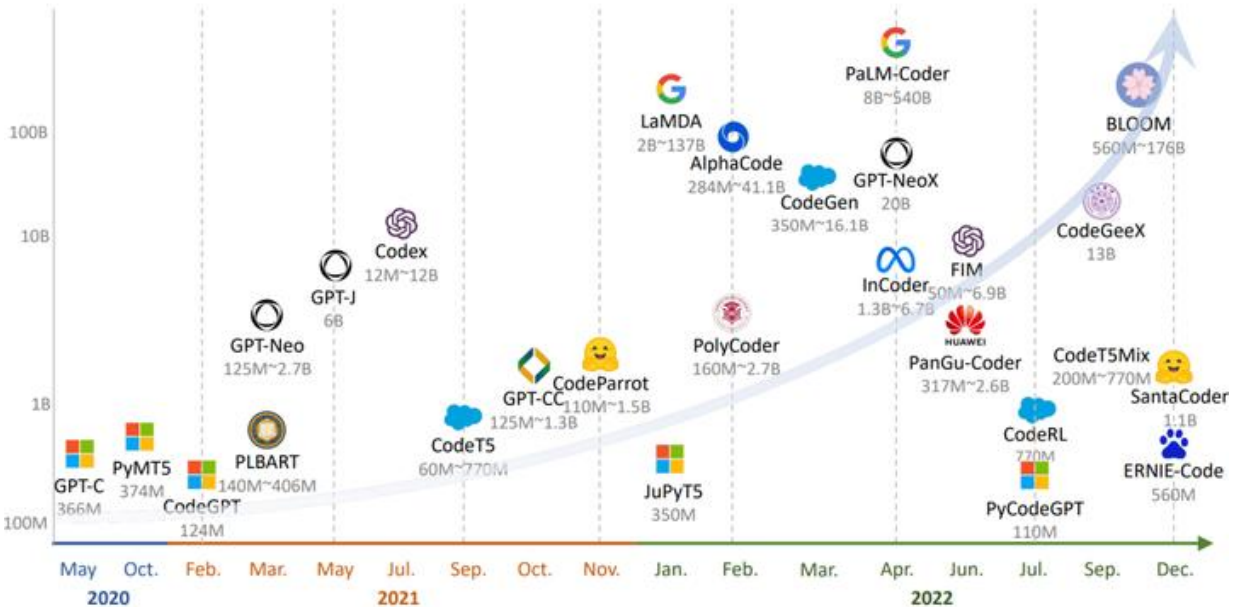
After clinical evaluation, she was diagnosed with seasonal allergic rhinitis.

Processing patient data comes with many limitations

- Commercial tools (Amazon, Google, Microsoft, other open source tools)
  - Cloud based: not ideal for patient sensitive data
  - Mainly built for English texts
  - Not tailored to our data distribution or clinical domain

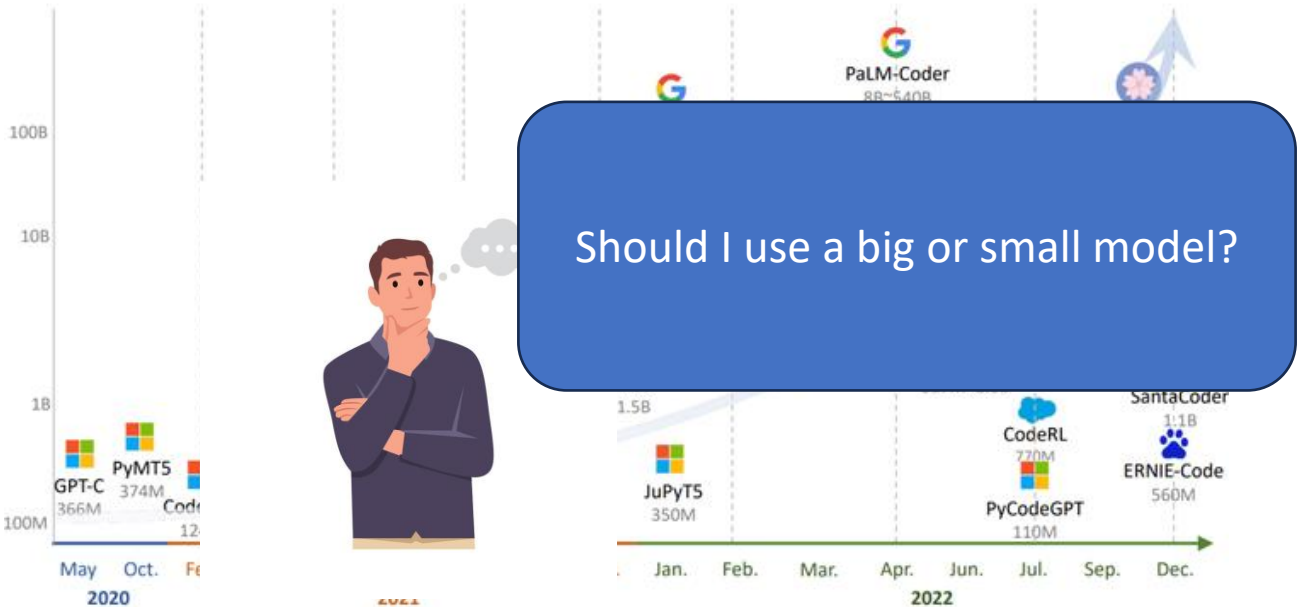
- Training our own NER system
  - Annotated data for German is scarce
  - Hospitals might not have a large GPUs

Large Language Models for NER



Source: "The Emergence of the Transformer - The Architectural Breakthrough, History, and the Future of The Model Behind ChatGPT"

Large Language Models for NER



Source: "*The Emergence of the Transformer - The Architectural Breakthrough, History, and the Future of The Model Behind ChatGPT*"

Large Language Models for NER

Research question:

Do we really need a large model to perform clinical NER in German or can smaller models like BERT do the job just as well?

Method

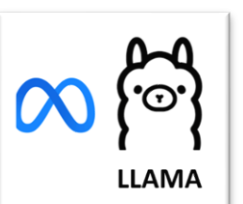
VS

- Compare a small LM **(G-BERT, 110M)** with a large LM **(Llama3.1, 8B)**
- Fine-tune them on two German NER datasets
- Evaluate
  - How do they perform with limited training data?
- Metric
  - entity-level F1 (exact match)

Models

model	architecture	# parameters	domain
deepset/gbert-base	Masked LM	110 M	General domain, pre-trained on German corpora
Llama 3.1	Generative LM	8 B	General domain (include German)

vs

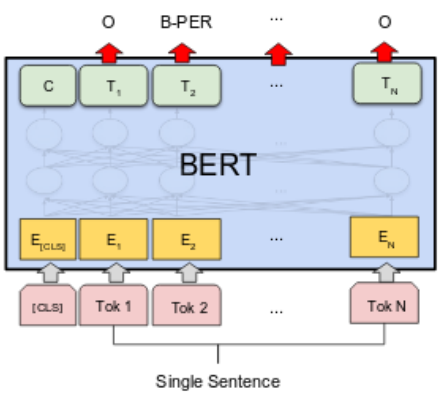


80 times larger than G-BERT

Fine-tuning process

G-BERT

Predict each token, IOB format



LLAMA

Instruct tuning

```
You are a german medical named entity recognition tagger.

Identify medical entities and types from the text. Only
give answer without explanation. These are possible entity
types: 'ACTIVEING'(active ingredient), 'DRUG', 'STRENGTH',
'FREQUENCY', 'DURATION', 'FORM'.

Please use only the given entity span and entity types.
Give the answer in the following format.

Answer:
Entity_Span1 (Entity_type1)
Entity_Span2 (Entity_type2)
Or return None if found nothing.

Text: Kalinor retard nach Kalium

Answer:
```

```
Kalinor retard (DRUG)
Kali (ACTIVEING)
```

Datasets

dataset	domain	language	# entity types
GGPONC 2.0	Oncology guidelines	German	7
CARDIO:DE	doctor's letters from the cardiovascular domain	German	6

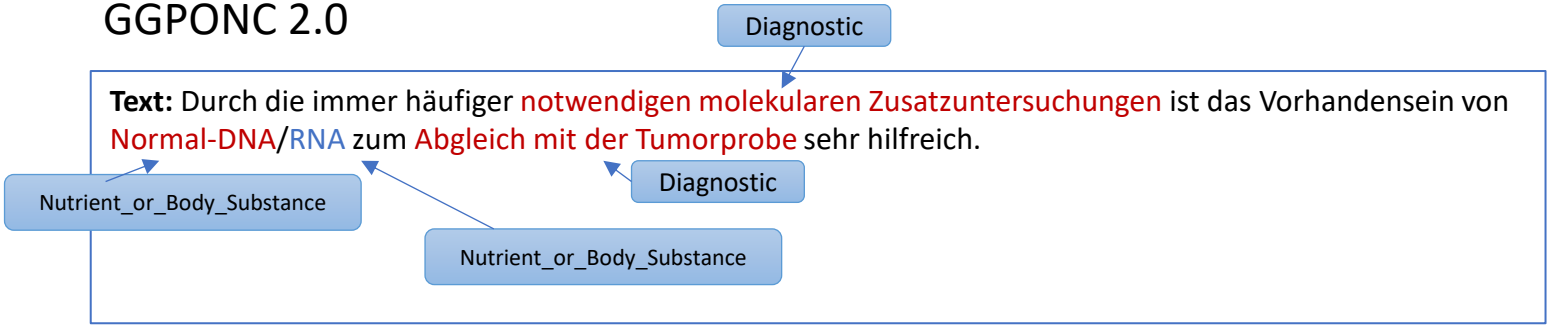
Entity types (GGPONC 2.0)

- Diagnosis_or_Pathology
- Clinical_Drug
- Diagnostic
- Other_Finding
- Therapeutic
- Nutrient_or_Body_Substance
- External_Substance

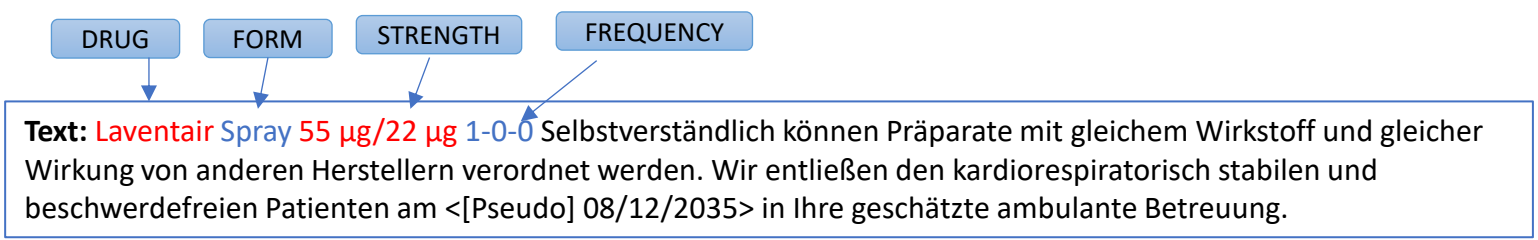
Entity types (CARDIO:DE)

- DRUG
- DURATION
- FORM
- FREQUENCY
- STRENGTH
- ACTIVEING

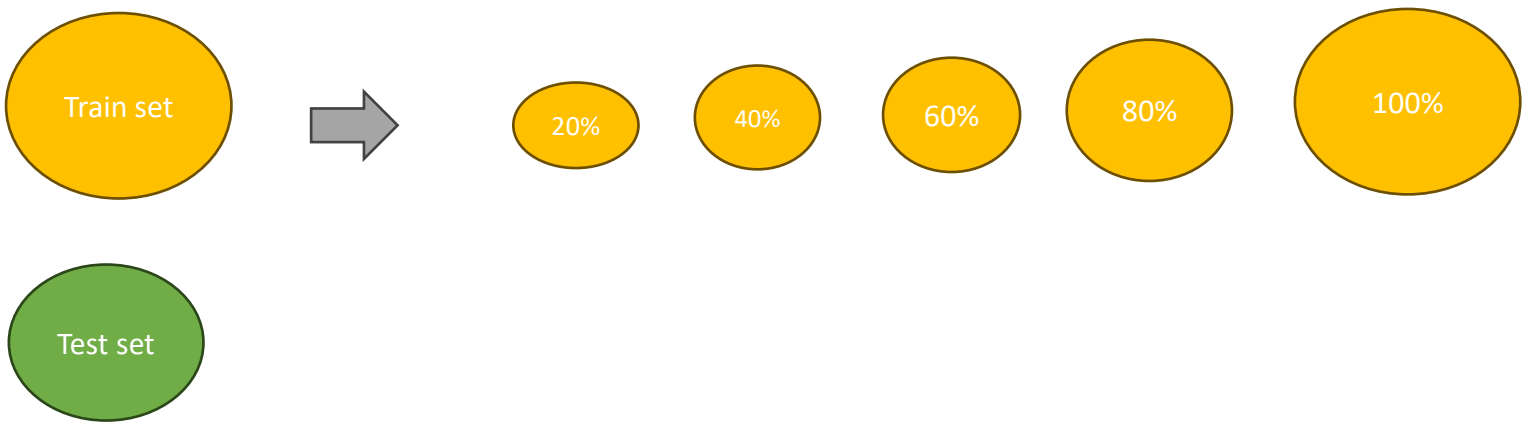
GGPONC 2.0

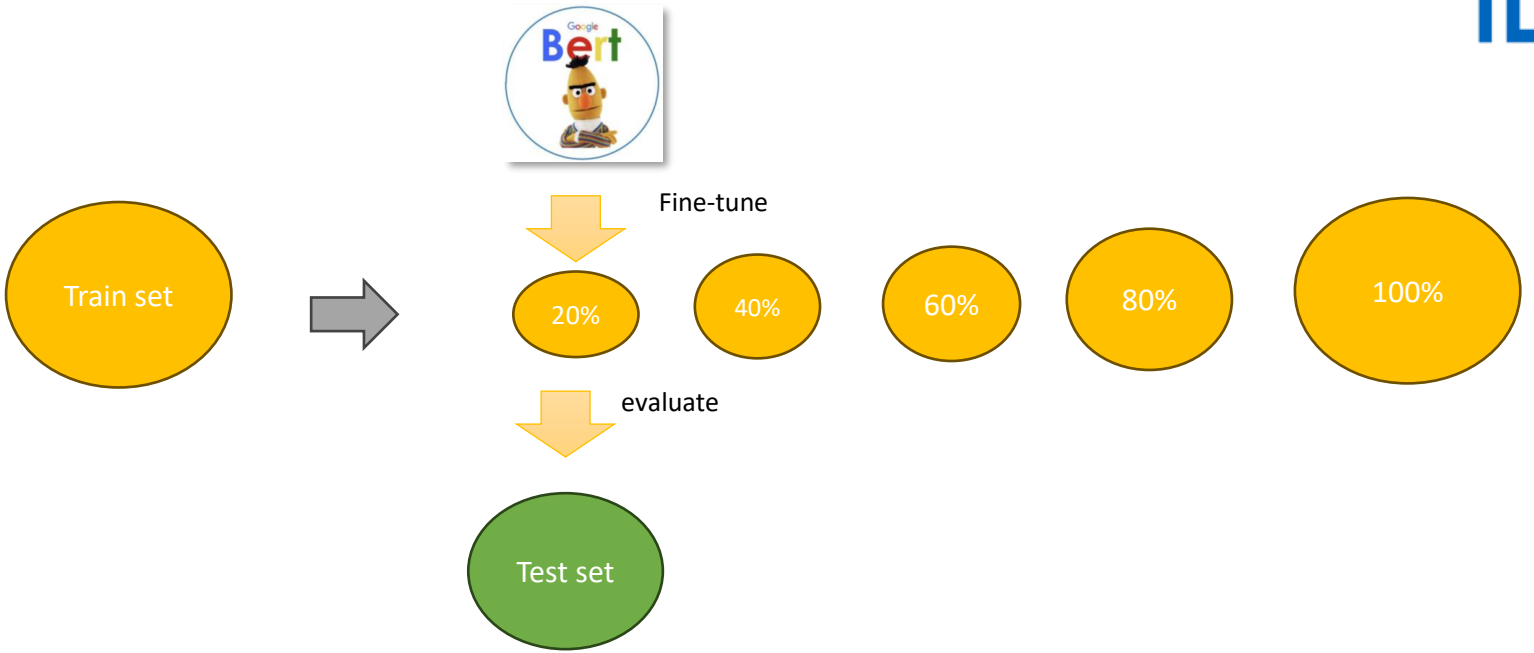


CARDIO:DE



Simulate low-resource settings




Bert
Train set
Fine-tune
20%
40%
60%
80%
100%
evaluate
Test set

Evaluation

Evaluation metric: F1-score entity level exact match

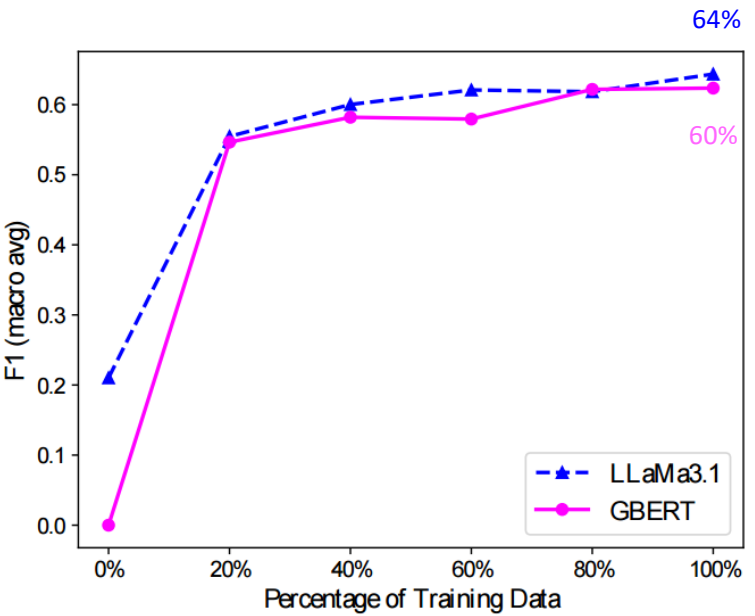
```
Insulin aspart 300 E. nach Plan Selbstverständlich können Präparate
```

Gold: ['Insulin aspart', 'ACTIVEING']

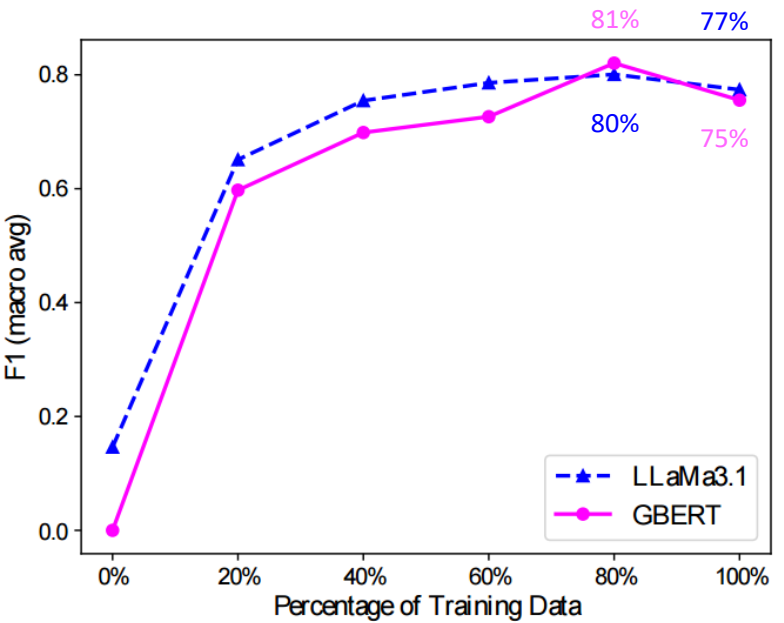
prediction	Entity level	Token level
['Insulin aspart', 'ACTIVEING']	1 correct	2 correct
['Insulin', 'ACTIVEING']	1 wrong	1 correct, 1 wrong

More data leads to better performance

GGPONC 2.0

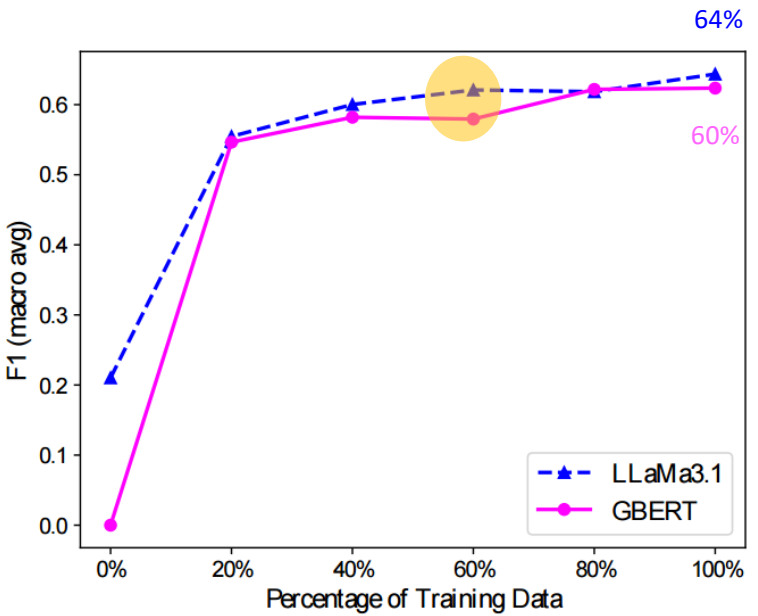


CARDIO:DE

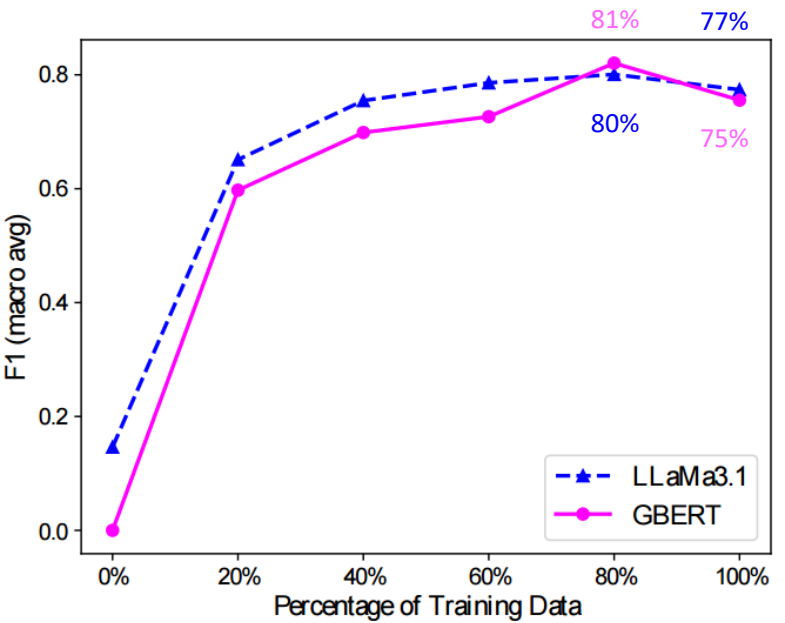


GGPONC - Similar performance

GGPONC 2.0

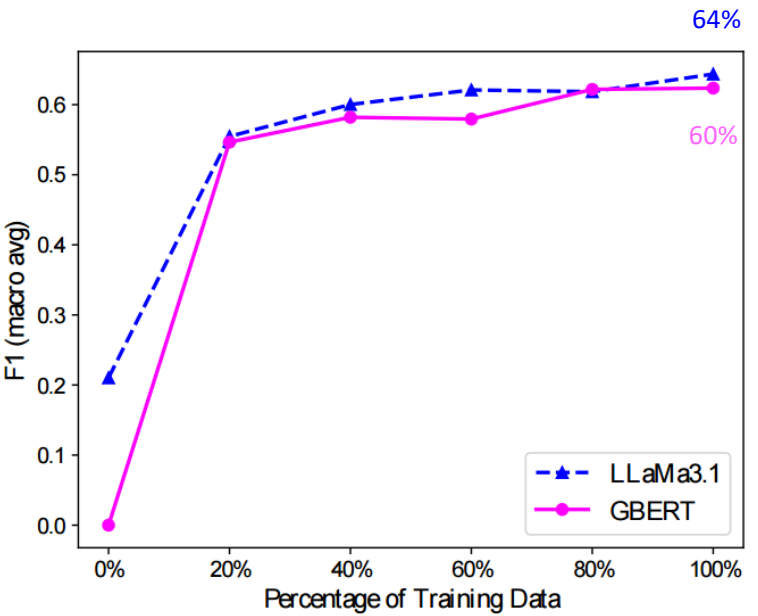


CARDIO:DE

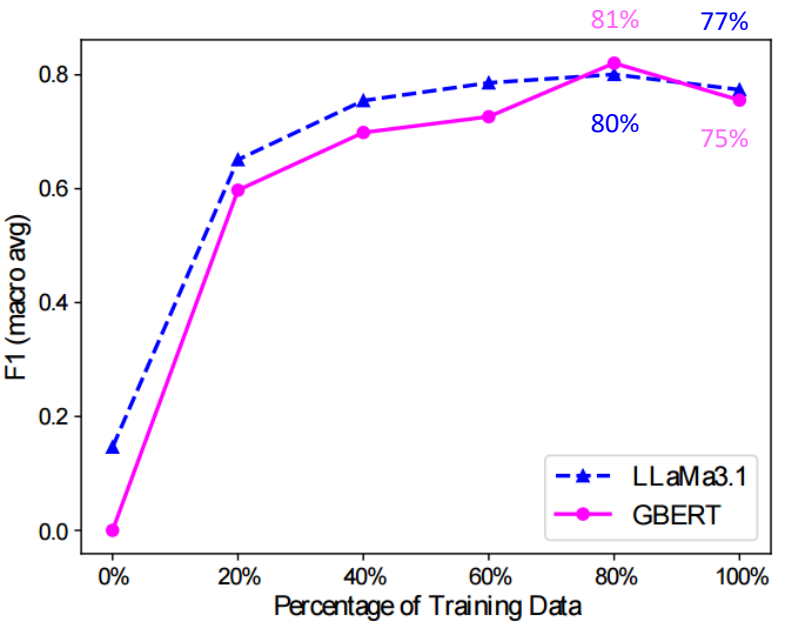


CARDIO:DE – LLaMa performs better in most scenarios

GGPONC 2.0

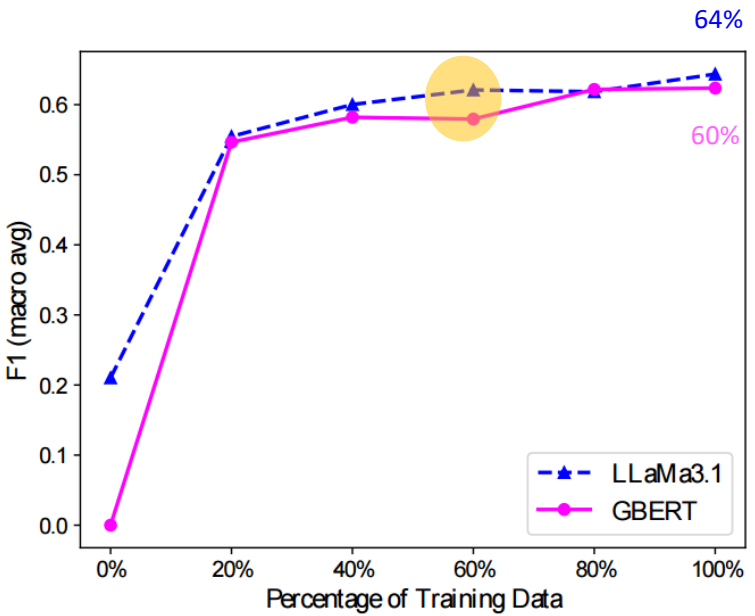


CARDIO:DE

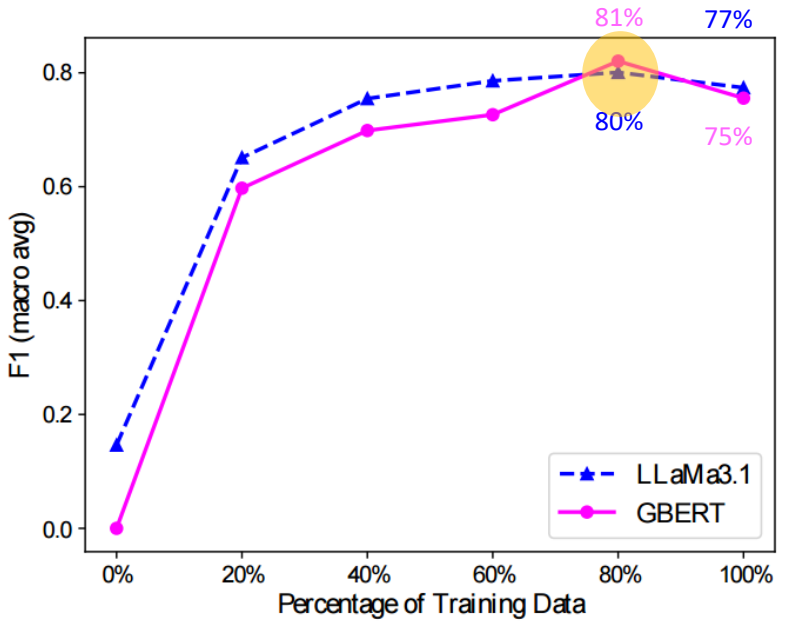


BERT could be more sensitive to outliers

GGPONC 2.0



CARDIO:DE



LlaMa suffers from hallucinations

- No hallucination

- Hallucinates span (2% out of all span errors), e.g., applying grammar corrections.
- Hallucinates entity type (0.2% out of type errors)
- But occurs less when trained with more data

BERT predicts more completely wrong span

Error type	BERT	Llama
Completely wrong span	~62%	~52%
Partially wrong span	~38%	~48%

`Insulin aspart 300 E. nach Plan Selbstverständlich können Präparate`

* Average over 2 datasets

BERT requires less computational resources and is faster

VS

- Use 16 GB GPU utilization during training
- 1 epoch, takes 30 minutes

- 80 times larger than BERT
- Use 48 GB GPU utilization during training
- 1 epoch, takes 2 hours
- Training is 4 times slower than BERT

Conclusion

- We compare a small MLM (G-BERT) and a large generative language model (Llama3.1) in NER, German clinical domain, focus on scenarios where training data is limited.

Results

- Both models can reach a very similar performance.
- When training data is limited Llama performs better.
- BERT seems to be sensitive to outliers, might overfit easily.
- Llama has the disadvantage that it hallucinates and requires more computational resource and time.

Takeaways

Do we really need a large model?

> It depends

- Small dataset –> use large generative models

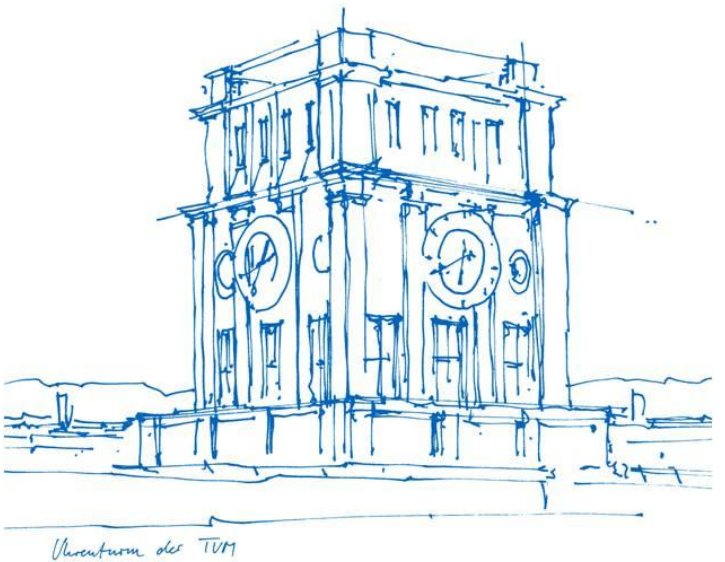
- Large dataset -> use small model like BERT
  - but make sure that annotation is high quality, because the model might be sensitive to outliers.

Thank you :)

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

suteera.seeha@tum.de

Linkedin: Suteera Seeha



Institute of AI and Informatics in Medicine (AIIM),
TUM University Hospital,
Technical University of Munich, Munich, Germany