Should you Mask 15% in Small Language Modelling?

Max Huppertz

Aaron Lingchi Chao

Katharina Strauss Soegaard

Timo Stephan Hermann Stoll

Team 7

2 | Abstract

Thesis: What is the impact of masking rates on small pretrained language models?

Conclusion: Masking rate impacts the performance of small-scale language models

Interesting Insights:

- Convergence and training speed is affected by masking rate
- A masking rate of 40% showed the best performance in downstream tasks

3 | Introduction

Why small models?

Large models achieve remarkable results (GPT-4 etc.) in NLP

But this comes at the cost of:

- High carbon footprint
- Huge energy and water consumption

→ Aim towards smaller and more efficient models

DistilBERT a distilled version of BERT: smaller, faster, cheaper and lighter, Sanh et al, https://arxiv.org/pdf/1910.01108.pdf
Energy and Policy Considerations for Deep Learning in NLP, Strubell et al., https://aclanthology.org/P19-1355.pdf
Data centre water consumption, David Mytton, https://www.nature.com/articles/s41545-021-00101-w

4 | Introduction

Why masking rate?

- Problem: **15%**, which is not universally optimal
- Larger models can benefit from higher masking rates

Masking rate affects:

- Performance on downstream tasks
- Sample efficiency
- Training time

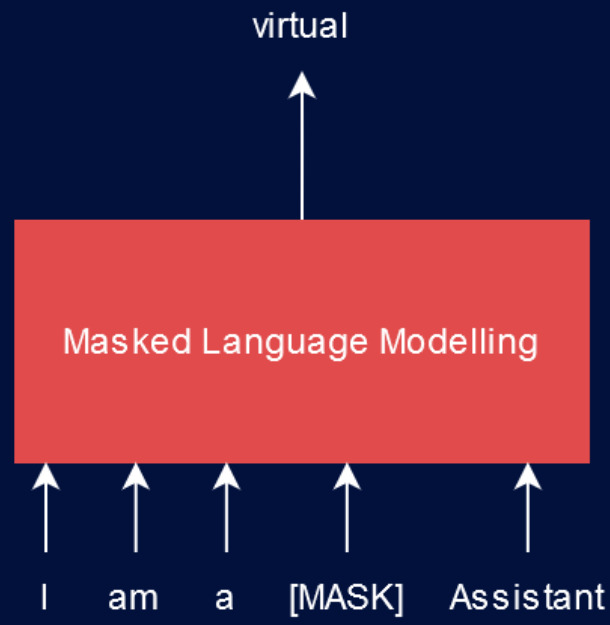
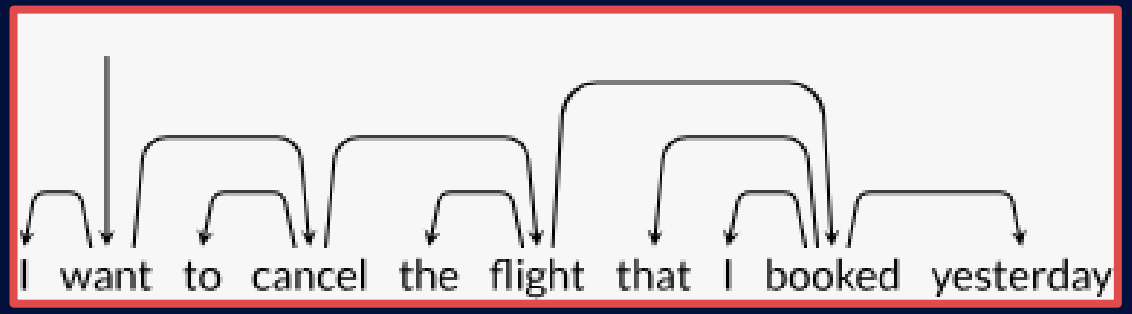


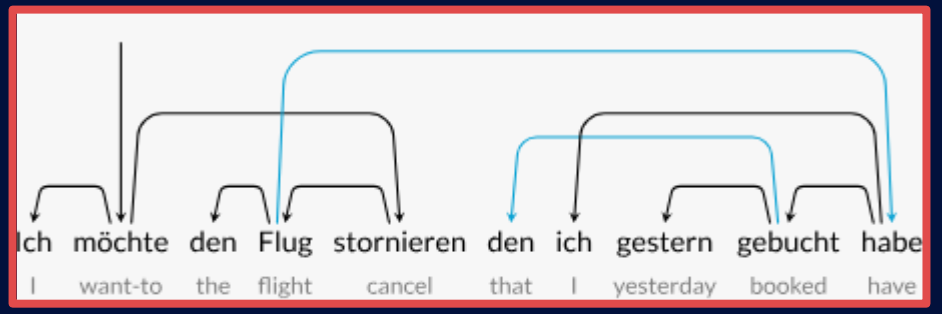
Fig: https://miro.medium.com/v2/resize:fit:880/1*0vMq-b5y1-XoA4TGJ3Akew.png
Should you Mask 15% in Masked Language Modeling?, Wettig et al., https://aclanthology.org/2023.eacl-main.217/

5 | Introduction

Why German Language?

- Non-Projective Syntax - *syntactic structures are not drawable without crossing 2 arcs*
- Different Grammar and Longer Words



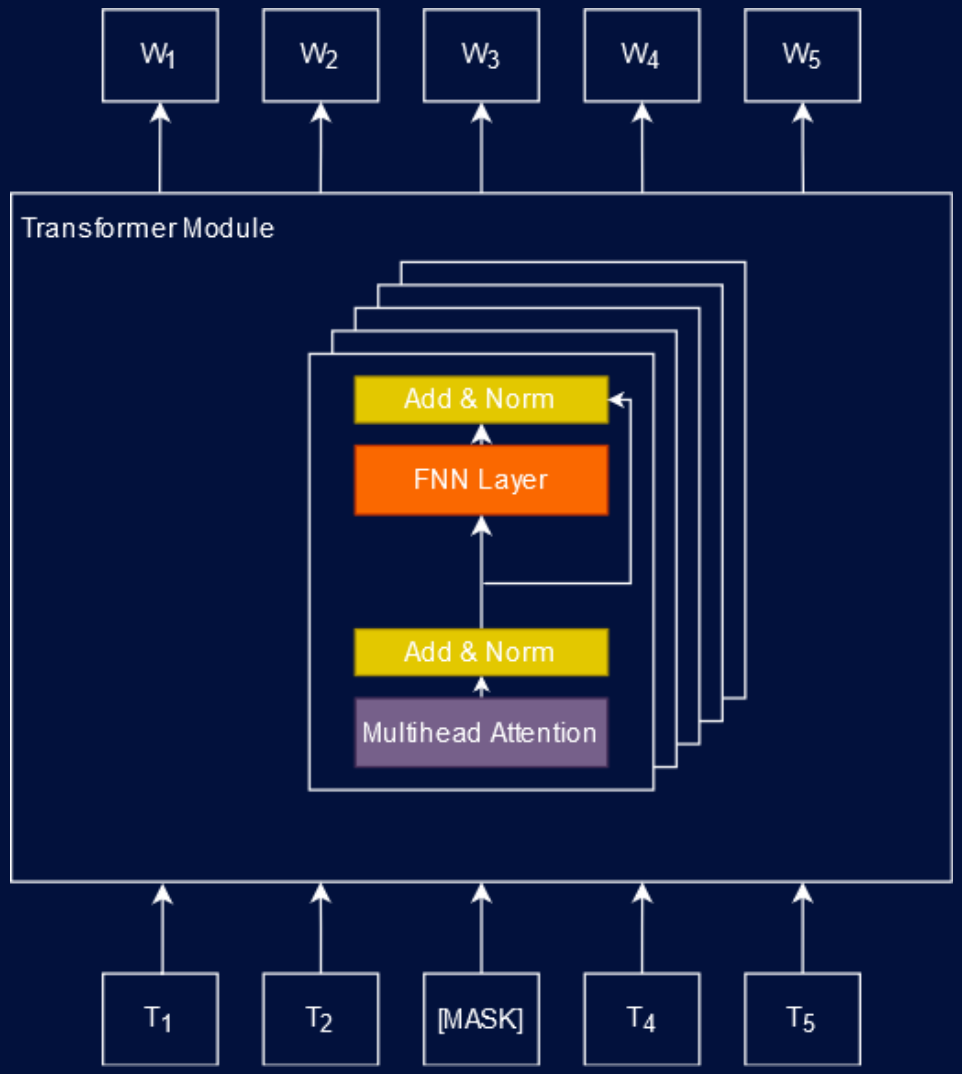


spaCy speaks German now, Wolfgang Seeker, https://explosion.ai/blog/german-model

6 | Approach

1	Replication of RoBERTa model given in baseline
2	Choice of German language datasets for training
3	Pretrain Bert Mini with varying masking rates
4	Evaluation on German benchmark datasets

7 | Methods



RoBERTa Model (Baseline):

- 12 Transformer Layers
- 768 neuron wide hidden layers
- 12 Attention heads

BERT Mini (Ours):

- 4 Transformer Layers
- 256 neuron wide hidden layers
- 4 Attention heads

Well read students learn better: On the importance of pretraining compact models, Turc et al., https://arxiv.org/pdf/1908.08962.pdf

8 | Methods


Huggingface
Transformer library
GermEval18
GermEval14
Gnad10k
Pretraining
Training Evauation
Finetuning
Evaluation
~24 hours
~1 hour
Masking rate
between 10% and 80%

9 | Datasets

Training

Wikipedia
German Data 2022

Evaluation

GermEval18
Coarse/Fine

GermEval14

Gnad10k

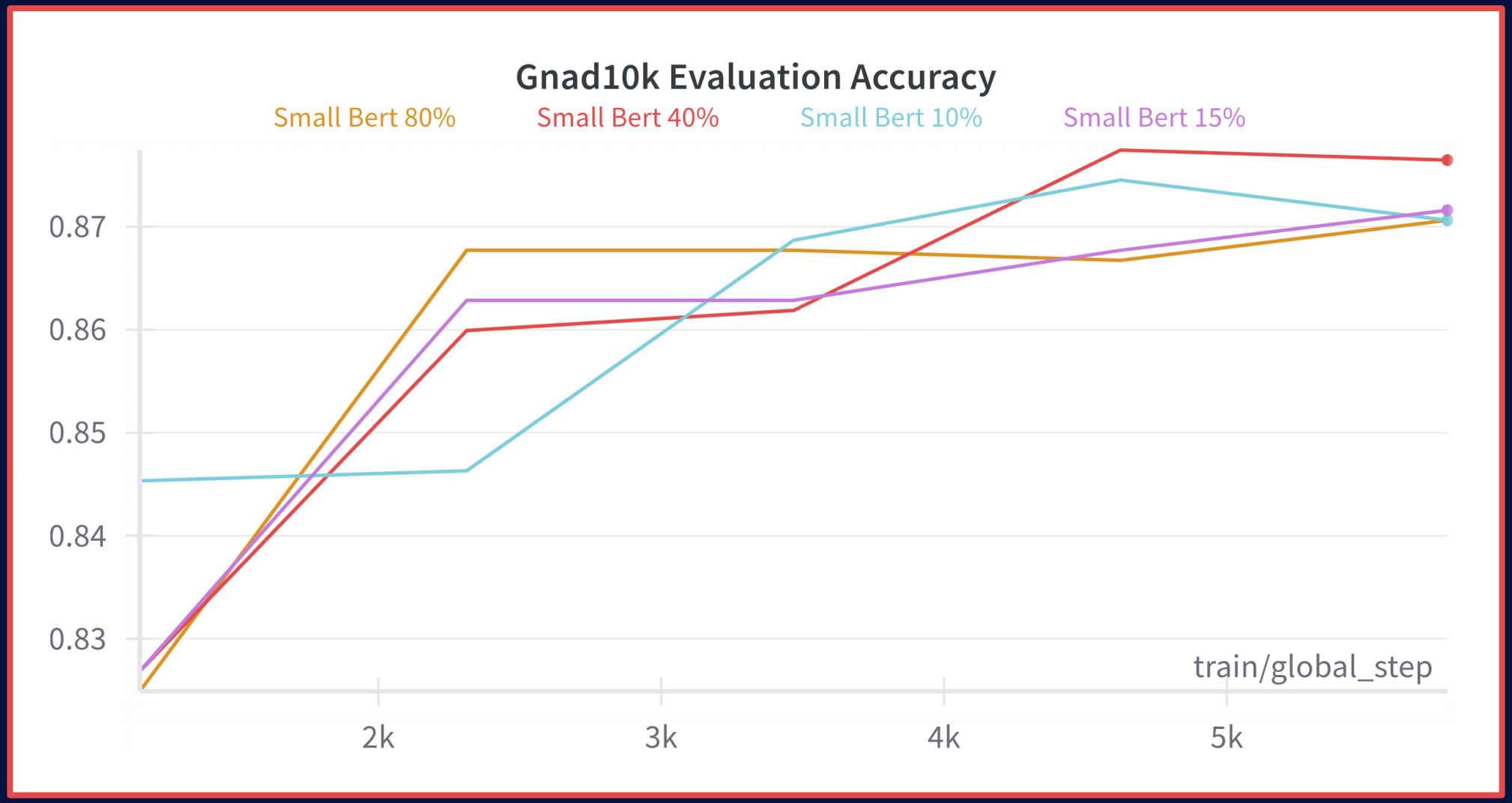
10 | Replication – GLUE

	QNLI	RTE	SST-2	MRPC	CoLA	STS-B
15%	**90.8**/ 90.9	**64.3**/ 67.3	**92.4**/ 93.3	**85.5**/ 77.0	**57.8**/ 59.2	**86.7**/ 87.7
40%	**91.2**/ 91.6	**62.1**/ 67.0	**92.3**/ 92.8	**87.5**/ 76.9	**59.3**/ 61.0	**86.7**/ 88.2
80%	**87.8**/ 87.9	**56.7**/ 58.6	**89.9**/ 90.5	**81.1**/ 72.1	**37.7**/ 38.7	**85.5**/ 86.3

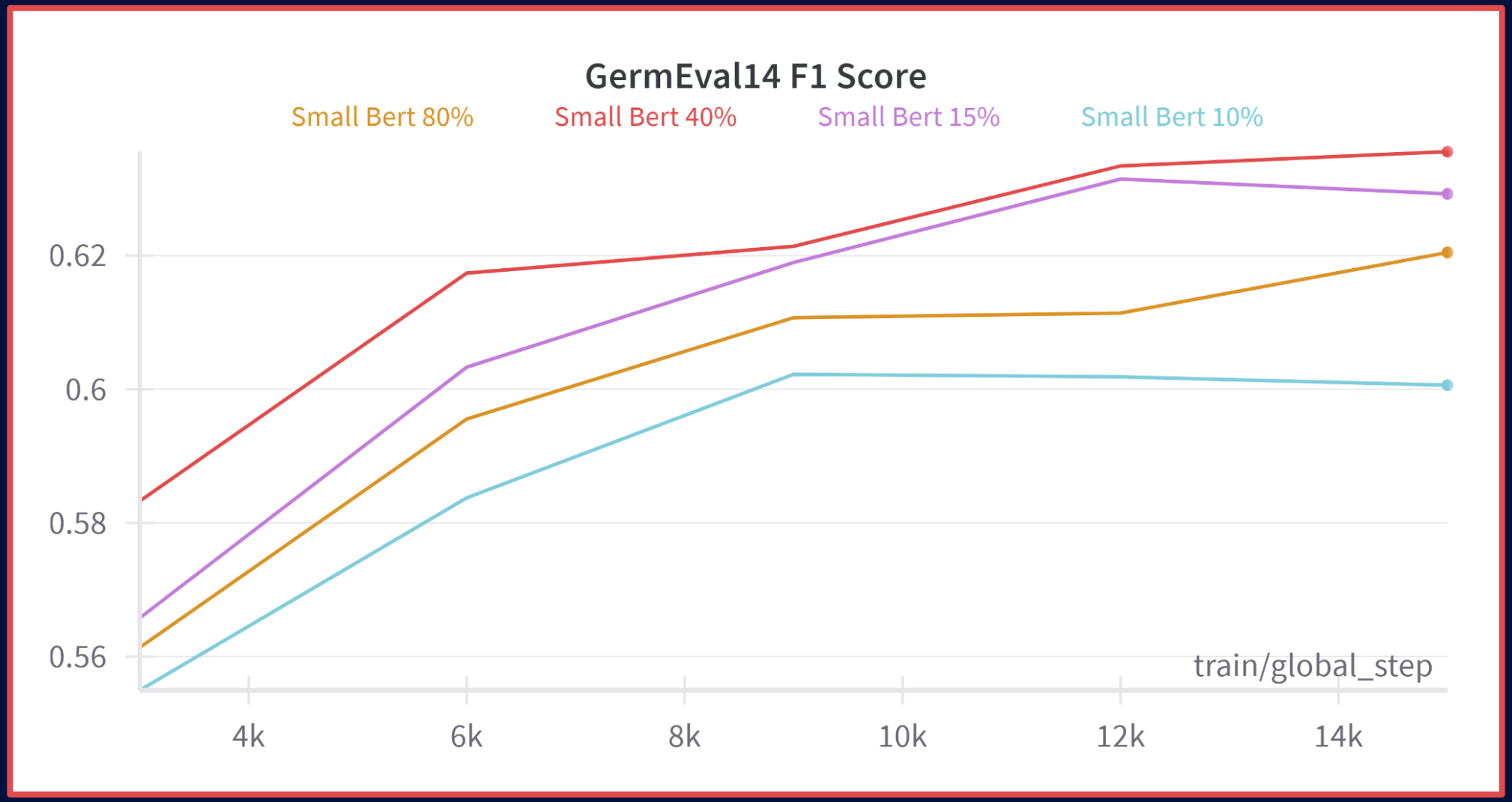
Our replication results are highlighted in light red

11 | Results – Training Loss

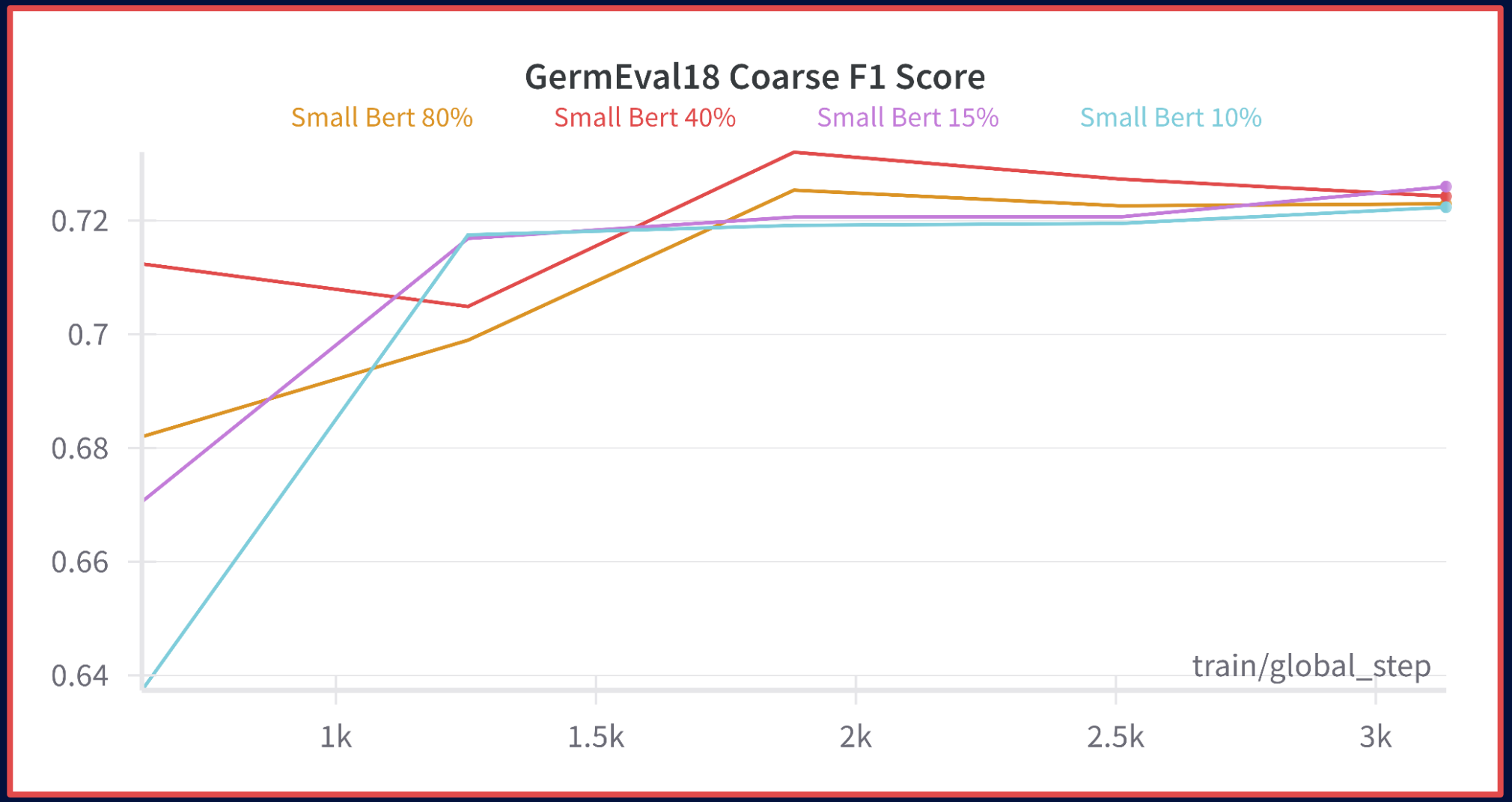
12 | Results – Gnad10k



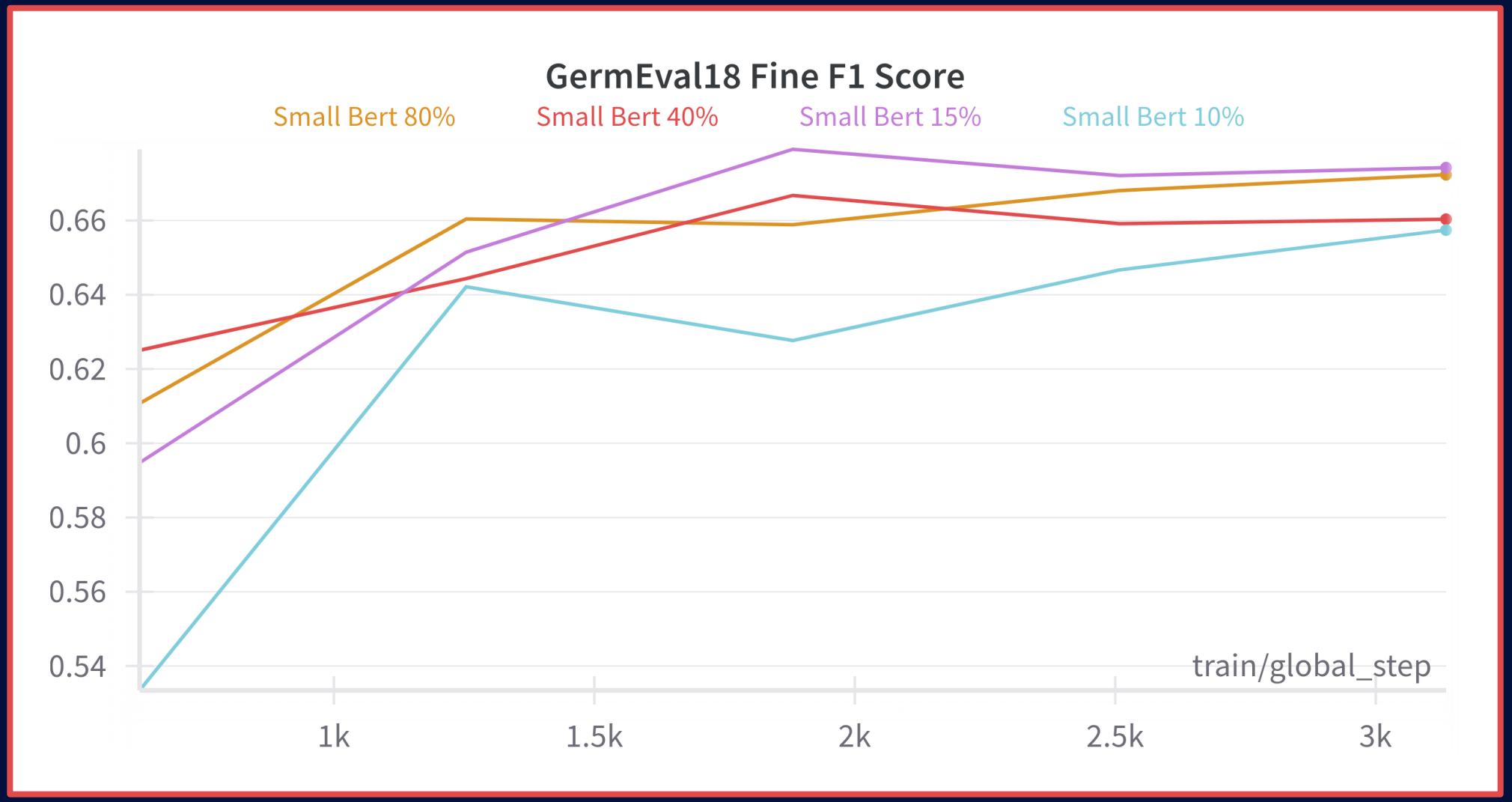
13 | Results – GermEval14



14 | Results – GermEval18 Coarse



15 | Results – GermEval18 Fine



16 | Results

	GermEval18 Coarse	GermEval18 Fine	GermEval14 (NER)	Gnad10k
Bert Mini 10%	0.7184 / **0.7205**	0.651 / 0.6404	0.6033 / 0.6579	0.8651
Bert Mini 15%	0.7206 / 0.7193	0.6621 / 0.6523	0.6273 / 0.6776	0.8667
Bert Mini 40%	**0.7223** / 0.7204	0.6585 / 0.6486	**0.6346** / **0.6845**	**0.8762**
Bert Mini 80%	0.719 / 0.7184	**0.6646** / **0.6537**	0.6204 / 0.6735	0.87082
DistilBERT	0.7054 / 0.7055	0.7652 / 0.7668	0.8270 / 0.8423	0.8901

Reported F1 score and Precision score for the evaluation datasets.
Accuracy is given for GNAD10k

 # Discussion – Results

- Higher masking rate does not reduce performance significantly, 40% performs best overall
- Convergence during pre-training is affected by masking rate
- Replication of Paper results was achieved for small scale model
- Masking rate appears to be independent of model size
- Although German differs significantly from English the results are similar

18 | Discussion – Limitations

- Evaluation was only conducted on the Bert Mini model, larger models left unexplored
- Training was only done on the Wikipedia dataset, larger models use bigger datasets
- Fewer available downstream task datasets in German with benchmark results
- Might want to treat the masking rate as a hyperparameter instead of setting it at 15%

19 | Conclusion

- Masking rate influences model performance: this also holds for small scale models
- A higher masking rate leads to slightly faster training and convergence
- Paper results were reproduced for small-scale models
- Limitations arise from the model size and available training data

Should you Mask 15% in Small Language Modelling?

Max Huppertz

Aaron Lingchi Chao

Katharina Strauss Soegaard

Timo Stephan Hermann Stoll

Team 7