Supplementary information

S1 Ablation test

S1.1 Comparing to medbertde

While *gbert-base-comb nocontext* was the overall best-performing model in our core experiments, the publicly available *medbertde-base nocontext* pretrained on medical data from scratch achieved superior results in frequent scenarios. Hence, we assessed, how *medbertde-base context* performs in comparison to our final model *gbert-large-comb context*. We trained a *medbertde-base context* model without further pretraining, as further pretraining did not show a consistent performance improvement in the core experiments. For 20 shots *medbertde-base context* achieved statistically significant better accuracy results than *gbert-large-comb context* (86.2% vs. 84.3%). Performance differences for 50 and 100 shots are not significant, while using 400 shots, *gbert-large-comb context* achieves better results (93.4% vs. 92.4%) (Suppl. Tab. S4).

Primary classes: With regard to the primary classes, the $F1$-score of *gbert-large-comb context* is significantly better than *medbertde-base context* for the *Anamnese* class with mean $+4.2$ percentage points. This support our hypothesis, that larger PLMs are superior on complex free text section classes (cf. Section 4). To a lesser extent, but significantly, *medbertde-base context* achieves better $F1$-scores for the *Medikation* class. (Suppl. Fig. S14)

S1.2 Inspecting `[SEP]` recognition

We observed significant performance drops for classes such as: *AllergienUnverträglichkeitenRisiken, Anrede* and *Mix*. To gain better understanding of this decline, (1) we performed *fine-grained class analysis* for samples from *Anrede* and (2) *analyzed Shapley values*. (1) We found that precision dropped from 98.2% to 48.1%. 99 out of 131 instances were misclassified as *Diagnosen*. Even if we use 400 training shots, the *gbert-base-comb context* model still achieves a low precision rate (56.8%). This can only be improved to 68.3% using a *gbert-large-comb context* model. Both precision scores are significantly below *nocontext* models with a precision of 98.2%.(2) Shapley values shed further light on typical patterns of this section samples. The three classes *AllergienUnverträglichkeitenRisiken, Anrede, Mix* typically contain only a single paragraph or sentence. *Anrede* paragraph are typically followed by a *Diagnosen* paragraph, containing section headers such as *Aktuelle Diagnosen:*.

Hence, to test the ability of PET models to recognize, that the sample to classify is between the two *[SEP]* token, we created nine artificial test samples by combining context paragraphs that are atypical for our dataset, as presented in Suppl. Fig. S15.

If we use a *gbert-base-comb context* model trained on 20 shots, the first sample in Suppl. Fig. S15 is still incorrectly classified with 97% as *Anrede*. In contrast, the second sample is correctly classified with 99% accuracy as *Medikation*.

Overall, 5/9 samples where still incorrectly classified as *Anrede* class.

We investigated another section class such as *AllergienUnverträglichkeitenRisiken*, which typically only contains a single paragraph, too, we observed a similar behaviour. Often the context models incorrectly classify samples from the previous section *Diagnosen* as *AllergienUnverträglichkeitenRisiken*. E.g. *Z.n. Bandscheibenvorfall 11.09.1941 [SEP] - Z.n. Hodentorsion 11.09.1941 [SEP] Kardiovaskuläre Risikofaktoren: Arterielle Hypertonie, Hypercholesterinämie, positive Familienanamnese, Nikotinanamnese: nie (English: History of disc prolapse on September 11, 1941 [SEP] - History of testicular torsion on September 11, 1941 [SEP] Cardiovascular risk factors: arterial hypertension, hypercholesterolemia, positive family history, smoking history: never.)*. These results raised the question: Are there often misclassifications at the first or final paragraph of a section class? The confusion matrix (Suppl. Fig. S16) shows that typical false positives involve such patterns. For example:

- Diagnosen often misclassified as *Anrede*
- *Diagnosen* and *Befunde* often misclassifed as *AllergienUnverträglichkeitenRisiken*
- *Medikation* and *Zusammenfassung* often misclassifed as *Abschluss*

This reveals, that contextualizing paragraphs can harm classification results for certain section classes. This is especially relevant for section classes, which usually contain single-paragraph samples. This suggests that in a few-shot learning scenario, smaller PLMs can have difficulty distinguishing testing instances from contexts, and hence do not sufficiently focus on the instances themselves.

S1.3 Removing section titles from data

We identified that our best performing model from the core experiments *gbert-base-comb nocontext* using 20 shots frequently misclassified samples containing section titles of our primary classes. 32% of the false negative samples of the *Medikation* class contained either the text sequence *Medikation bei Aufnahme: (English: Medication on admission)* or *Medikation bei Entlassung: (English: Medication at discharge)*. A similar classification error we observed for the *Anamnese* class: 81% of false negatives contain the text sequence *Anamnese*. While in the training samples for *Medikation* we did not identify any section titles, there was a single title *Anamnese:* in the training set of *Anamnese*.

Adding context and increasing model size could significantly avoid these kind of errors, still 5% of the false negatives of the *Medikation* class of our final model *gbert-large-comb context* contained these kind of text sequences. Hence, we trained our final model *gbert-large-comb context* on a modified training and test set, filtered by a list of the most common section titles (Suppl. Fig. S17). In Suppl. Tab. S5 shows, that accuracy could be increased over all few-shot sizes by approximately 2%. Suppl. Fig. S18 shows, that both primary classes can improve F1-scores for 20 and 50 shots. In contrast, the models trained on the full training set, slightly decrease in performance. This is not surprising, as in contrast to the few-shot sets, the full training set frequently contains section titles.
However, it is important to note that these results can not be compared directly to the experimental results with included section titles, since we modified the training and test data set. Considering experimental limitations in clinical routine, it may be beneficial to avoid the use of section titles as they can be often well identified through manual patterns and heuristics. This approach is especially relevant if only smaller PLMs are employed with strong sequence length restrictions due to limited resources.

S1.4 Classifying nocontext samples using a context model

In Suppl. Fig. S19 we show Shapley values of a sample without further context with the gold label *Zusammenfassung* classified by (a) *gbert-base-comb context* and (b) the *gbert-large-comb context* model. *Gbert-base-comb context* shows very similar token contributions with respect to *Zusammenfassung* as *gbert-base-comb nocontext* in S8a. But both base models incorrectly classify the sample.
In contrast, *gbert-large-comb context*, correctly assigns the *Zusammenfassung* class with a probability of 79%. Adding context paragraphs increases this to 99% (see S13b). Interestingly, most of the input token positively contribute to the correct class, with the exception of *Aufnahme*. This is expected, as this token frequently negatively contributed to *Zusammenfassung* in various experimental setups.

S2 Baseline - support vector machine

While deep learning methods became the state-of-the-art for text classification tasks, statistical machine learning approaches such as Support Vector Machines (SVM) remain highly prevalent (Pomares-Quimbaya et al. 2019). Therefore we trained a SVM for our core experimental setup to compare its performance to our neural SC baseline and our best performing PET approach in the core experiments *gbert-base-comb nocontext* (Figure S1).
PET is always outperforming both SVM and SC for all shot sizes. Only if shot size is ≤ 20 the SVM outperforms our neural SC baseline model. If shot sizes are ≥ 50 SC and PET consistently outperform the SVM. We used the LinearSVC implementation and TfidfVectorizer for text encoding, both with default hyperparameters, as implemented in scikit-learn version 1.0.2. (Buitinck et al. 2013).

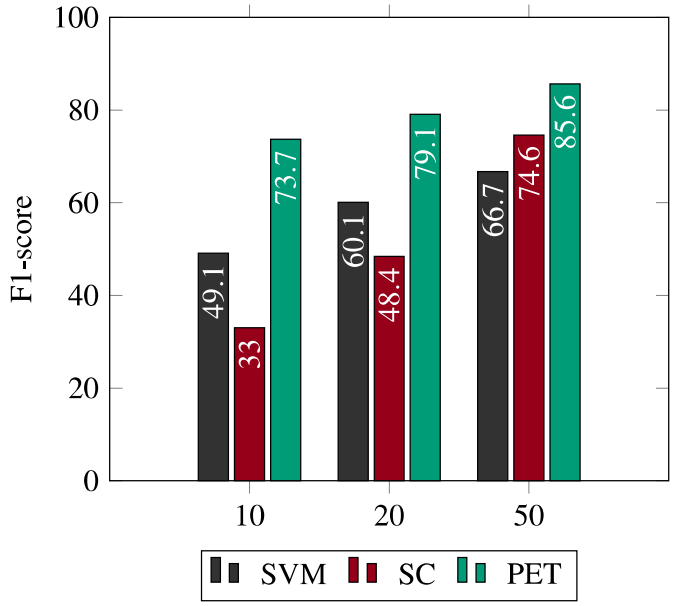
S3 Hyperparameters

Further pretraining: We applied the following hyperparameters for pretraining experiments described in Section 2.2: vocabulary size: 30,000; maximum sequence length: 512;

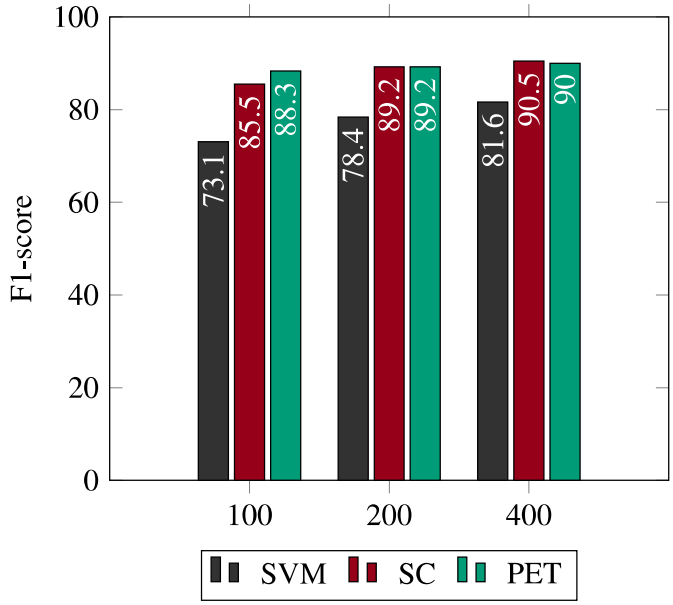
(1) task-adaptation:
- data: CARDIO:DE corpus
- epochs: 100, batch size: 24, fp16: True, gradient accumulation steps: 4
- 1×RTX6000 graphics processing unit (GPU) with 24 GB video random access memory (VRAM)
- Training time: $\sim$ 2h

(2) domain-adaptation:
- data: 179,000 German doctor's letters + GGPONC
- epochs: 3, batch size: 16, fp16: True, gradient accumulation steps: 1
- 2×RTX6000 GPUs with each 24 GB VRAM
- Training time: $\sim$ 17h

(3) combined:
- data: CARDIO:DE corpus
- epochs: epochs: 100, batch size: 24, fp16: True, gradient accumulation steps: 4
- 1×RTX6000 GPU with 24 GB VRAM
- Training time: $\sim$ 2h



(a) Comparing SVM, SC and PET using few-shot sets: 10, 20, 50



(b) Comparing SVM, SC and PET using few-shot sets: 10, 20, 50

Figure S1: **Baseline comparison SVM, SC and PET:** Comparing model performance using core experimental setup. Comparing Support Vector Machine (SVM), BERT with a sequence classification head (SC) and PET.

PET and SC experiments:

- All PET experiments were conducted on a single NVIDIA A40 GPU with 40 GB VRAM. However, we also conducted PET experiments on NVIDIA P4 with 8 GB VRAM using BERT-base models by only reducing evaluation batch size at inference time.
- Hyperparameters PET and SC: BERT-base models: training batch size 4, evaluation batch size: 64; BERT-large models: training batch size 4, evaluation batch size 16.

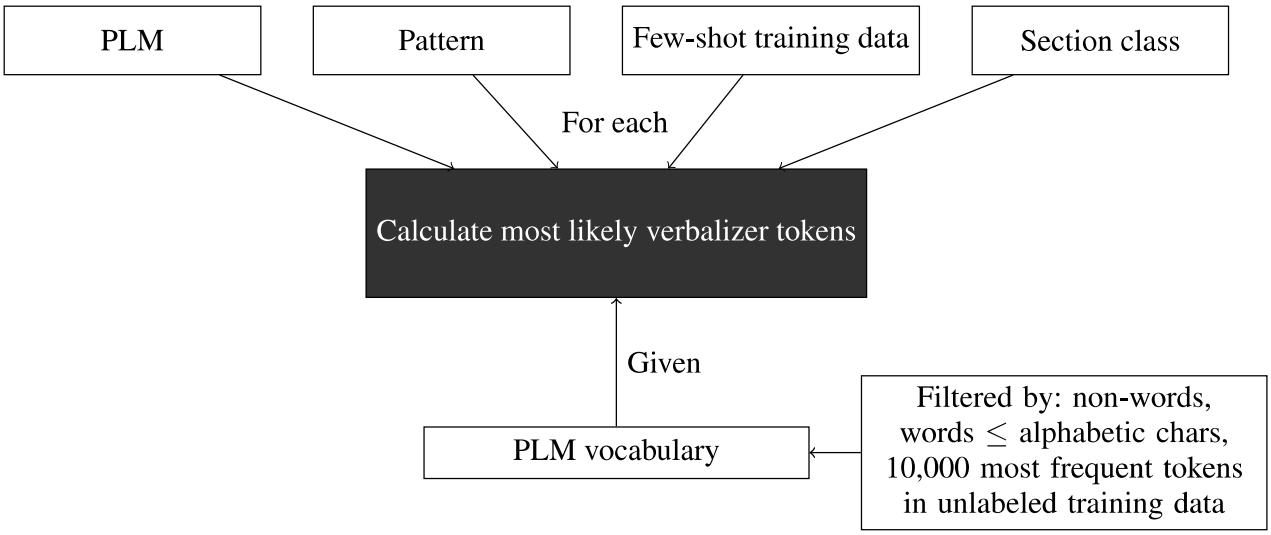


Figure S2: The PETAL workflow: PETAL calculates the most likely verbalizer token per label for each (1) PLM, (2) prompt pattern, (3) few-shot training set. The verbalizer token must be part of the PLM's vocabulary.

```
| - 10shots/
|   | - set_1.csv
|   | - set_2.csv
|   | - set_3.csv
|   | - unlabeled_1.csv
|   | - unlabeled_2.csv
|   | - unlabeled_3.csv
| - holdout/
|   | - full_holdout.csv
```

Figure S3: Few-shot data: Example folder structure for the 10shot data set including the heldout data set.

- Each experiment conducted with three different training sets and two random seeds (in total six setups). To increase comparability, we always selected models trained on training set 3 and with random seed 123 to investigate Shapley values.

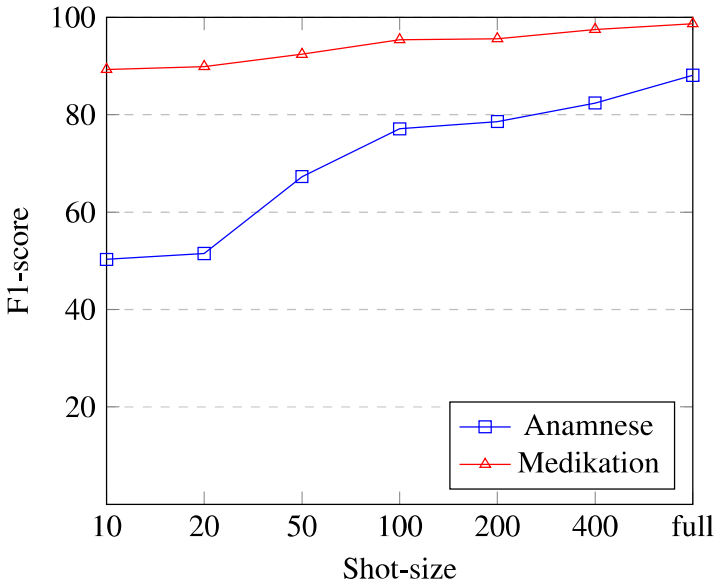


Figure S4: **Core experiments: Primary class $F1$-score for all shot sizes**. $F1$-score per few-shot sizes for primary classes with no context using *gbert-base-comb nocontext*.

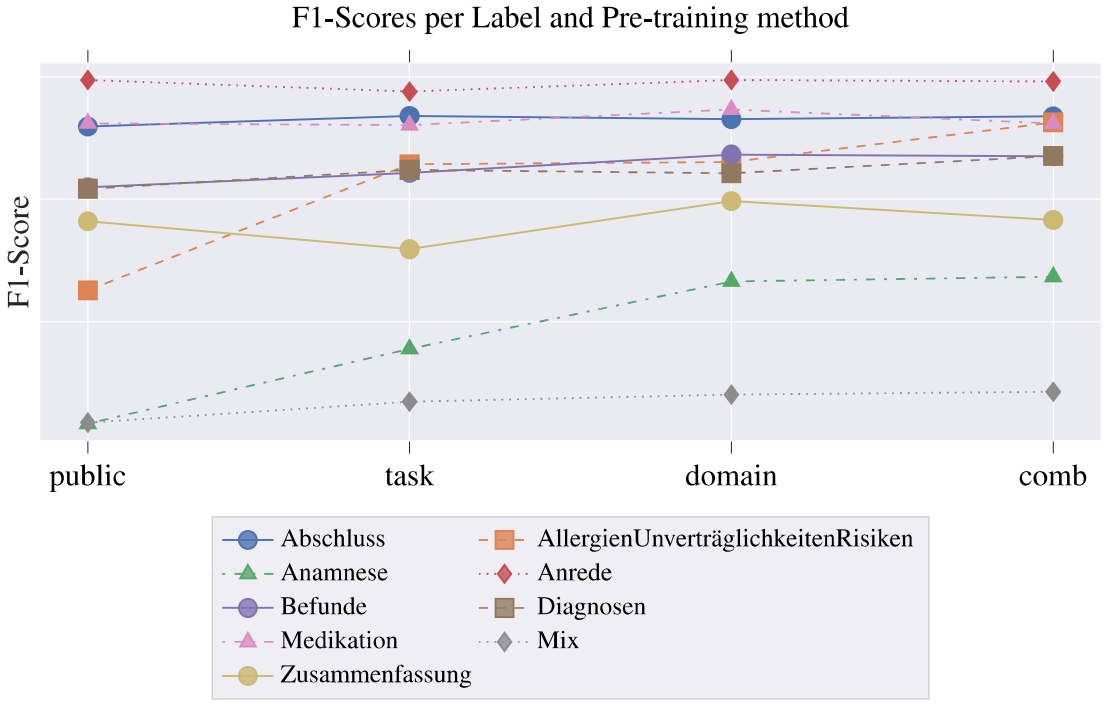


Figure S5: F1-scores per label per pretraining method using *gbert-base nocontext*

.

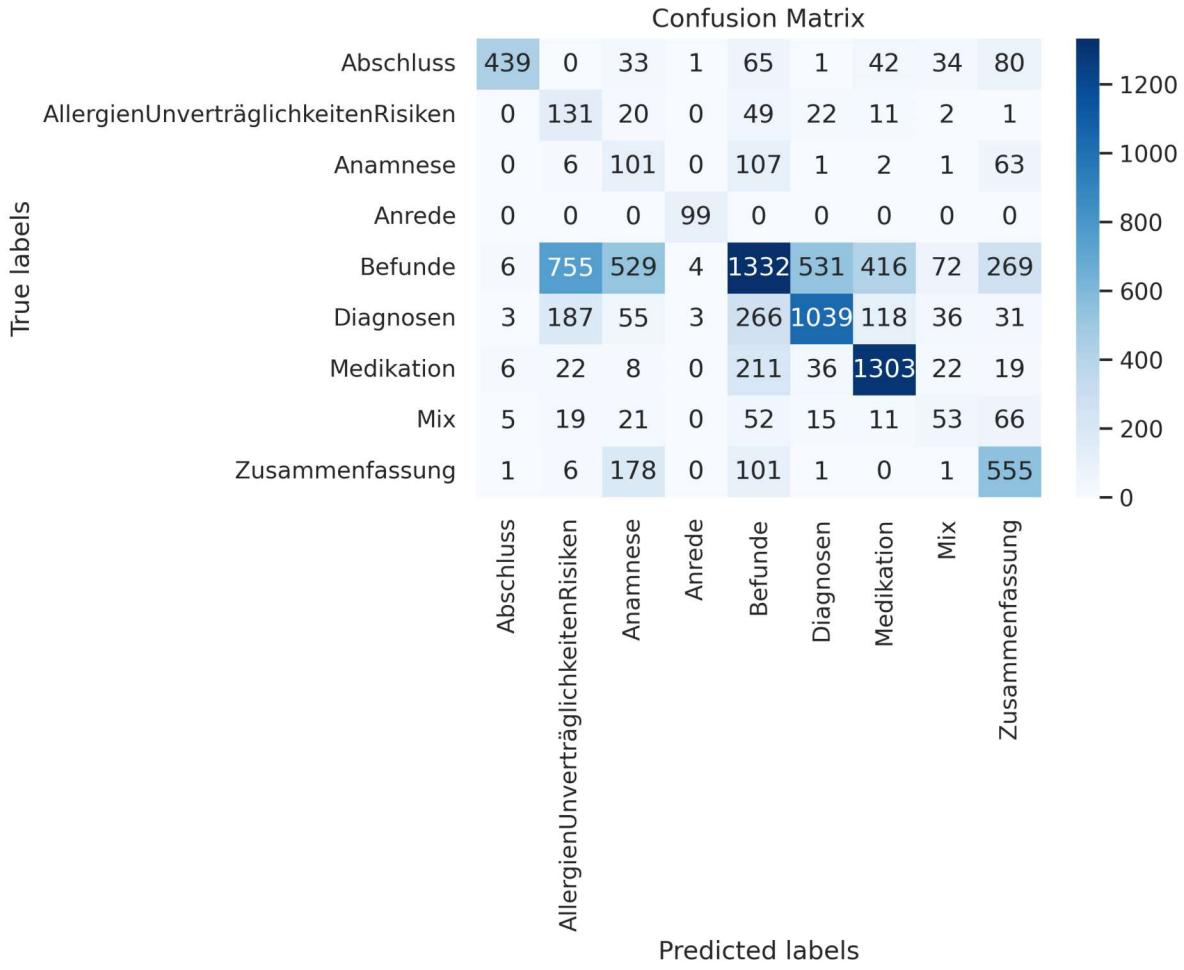


Figure S6: Confusion matrix for *gbert-base* trained on 20 shots on training set 3 with initial seed 123.

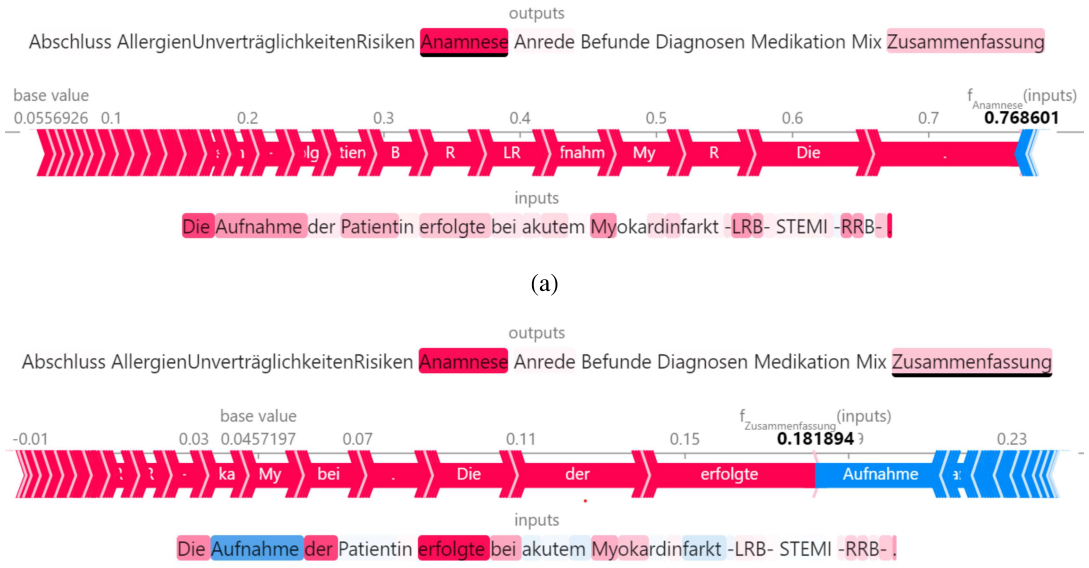


Figure S7: Shapley values for *gbert-base-comb nocontext* for predicted class comparing (a) *Anamnese* and (b) *Zusammenfassung* using 20 training shots. Shapley values with respect to predicted label (underlined). Shapley values per sub tokens. Legend: Red: positive contribution, Blue: negative contribution.

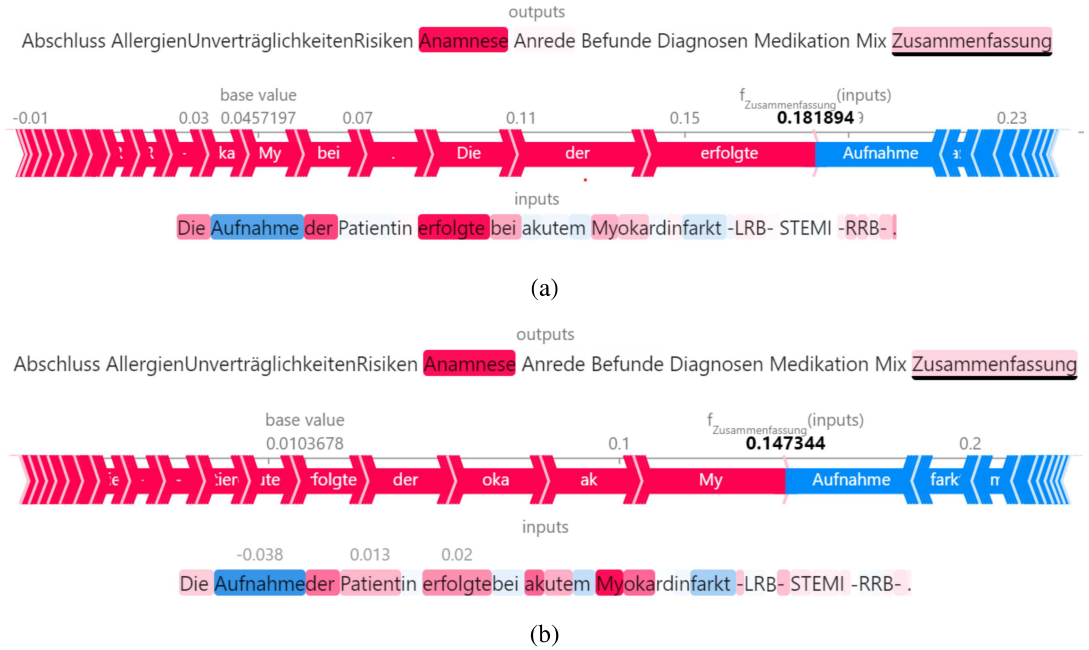


Figure S8: Shapley values for predicted class *Zusammenfassung* comparing (a) *gbert-base-comb nocontext* and (b) *gbert-large-comb nocontext* with 20 training shots. Shapley values with respect to predicted label (underlined). Shapley values per sub tokens.
Legend: Red: positive contribution, Blue: negative contribution.

	Training set	Test set
Anrede (Salutation/Greeting)	402	99
AktuellDiagnosen (Current Diagnosis)	3,298	694
Diagnosen (Diagnosis)	4,725	1,044
AllergienUnverträglichkeitenRisiken (AllergiesIntolerancesRisks)	1,031	236
Anamnese (Patient Medical History)	1,188	281
AufnahmeMedikation (Admission Medication)	2,058	593
KUBefunde (Body Findings)	4,194	1,105
Befunde (Findings)	9,636	2,519
EchoBefunde (Echocardiogram Findings)	1,566	290
Labor (Laboratory)	55,420	12,220
Zusammenfassung (Summary)	3,645	843
Mix (Mix)	945	242
EntlassMedikation (Discharge Medication)	4,090	1,034
Abschluss (Closing Remarks)	2,805	695
Total	95,003	21,895

Table S1. : Number of samples per section class per CARDIO:DE corpus split. English translations in round brackets.

PLM	Pretrained	Method	Few-shots
gbert-base	public	PET&SC	10, 20, 50, 100, 200, 400
	task		
	domain		
	comb		
medbertde-base	public		
	task		
	domain		
	comb		

Table S2. : Setup for core experiments: Experimental overview for our core experiments including PLMs, pretraining method, learning method and few-shot sizes.

Shot size	All templates	Null prompt templates
20	79.1	78.2
50	85.6	84.6
100	88.3	88.5
400	89.7	90
full (SC)	96.7	

Table S3. : **Null prompts:** accuracy scores for *gbert-base-comb nocontext* PLMs using all templates or null prompts on four few-shot sizes.

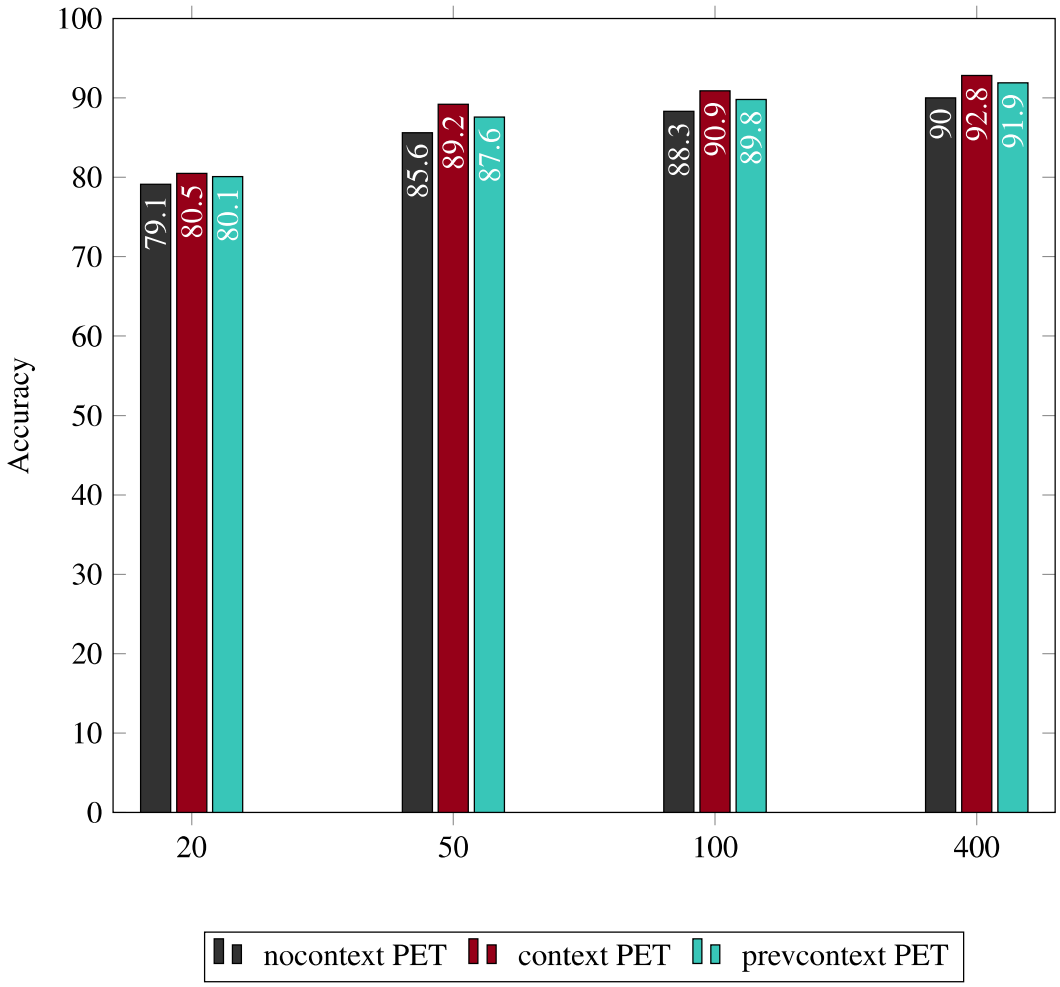


Figure S9: **Context types**: accuracy scores for different context types: (1) no context, (2) context, (3) prevcontext and few-shot sizes 20, 50, 100 and 400 using PET.

Shot size	gbert-large-comb	medbertde-base
20	84.3	86.2
50	89.4	90.3
100	91.3	91.3
400	93.4	92.4

Table S4. : Comparing *gbert-large-comb context* and *medbertde-base context* trained with all templates.

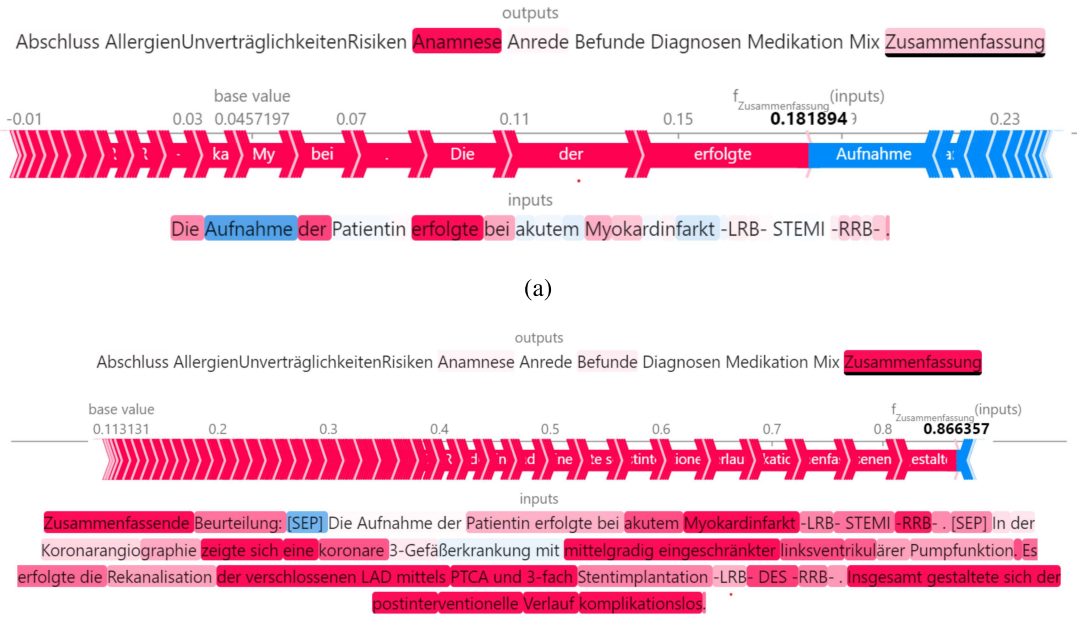


Figure S10: Shapley values for *gbert-base-comb context* for predicted class Zusammenfassung comparing (a) *gbert-base nocontext* and (b) *gbert-base context* with 20 training shots. Shapley values with respect to predicted label (underlined). Shapley values per sub tokens.
Legend: Red: positive contribution, Blue: negative contribution.

Shot size	PET including section titles	PET excl. section titles
20	84.3	86.7
50	89.4	91.1
100	91.3	93.4
400	93.4	95.8

Table S5. : F1-score results using *gbert-large-comb context* trained with and without section titles in training and test data.

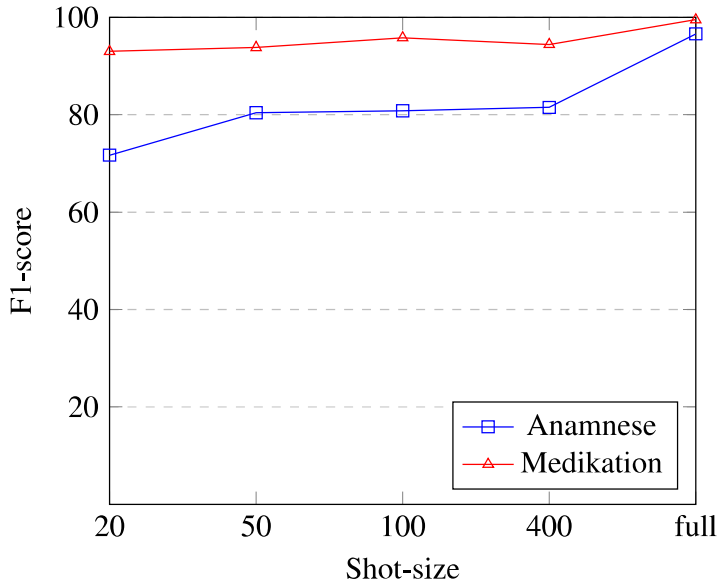


Figure S11: **Additional experiments: Primary class $F1$-score for all shot sizes**: Accuracy scores per few-shot sizes for primary classes using *gbert-large-comb context*.

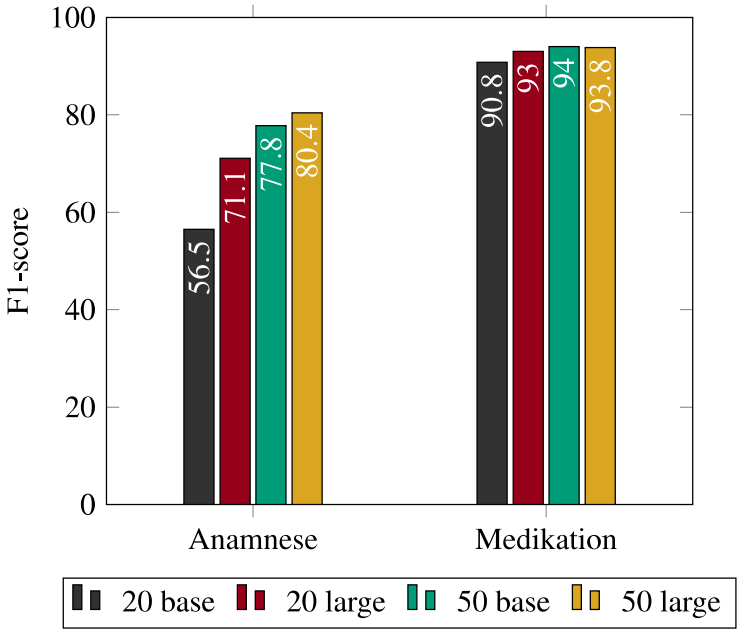


Figure S12: Combining best performing methods: comparing accuracy scores for *gbert-large-comb context* vs. *gbert-base-comb context* with all templates on two few-shot sizes for primary classes.

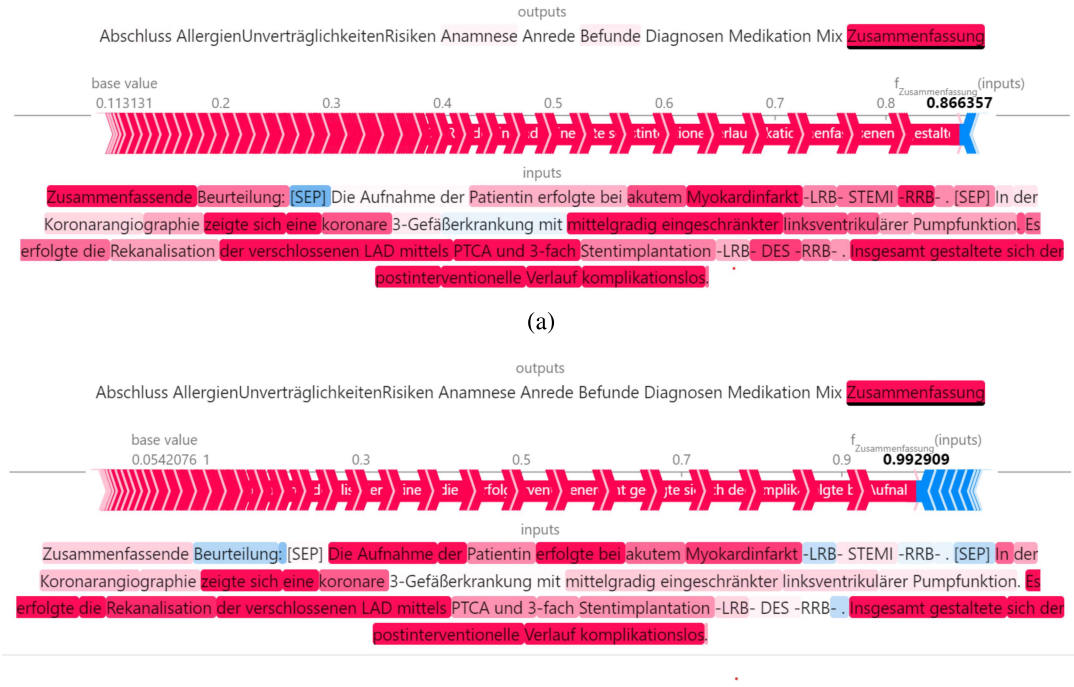


Figure S13: Shapley values for gbert-large-comb, context for predicted class *Zusammenfassung* comparing (a) *gbert-base context* and (b) *gbert-large context* with 20 training shots. Shapley values with respect to predicted label (underlined). Shapley values per sub tokens.
Legend: Red: positive contribution, Blue: negative contribution.

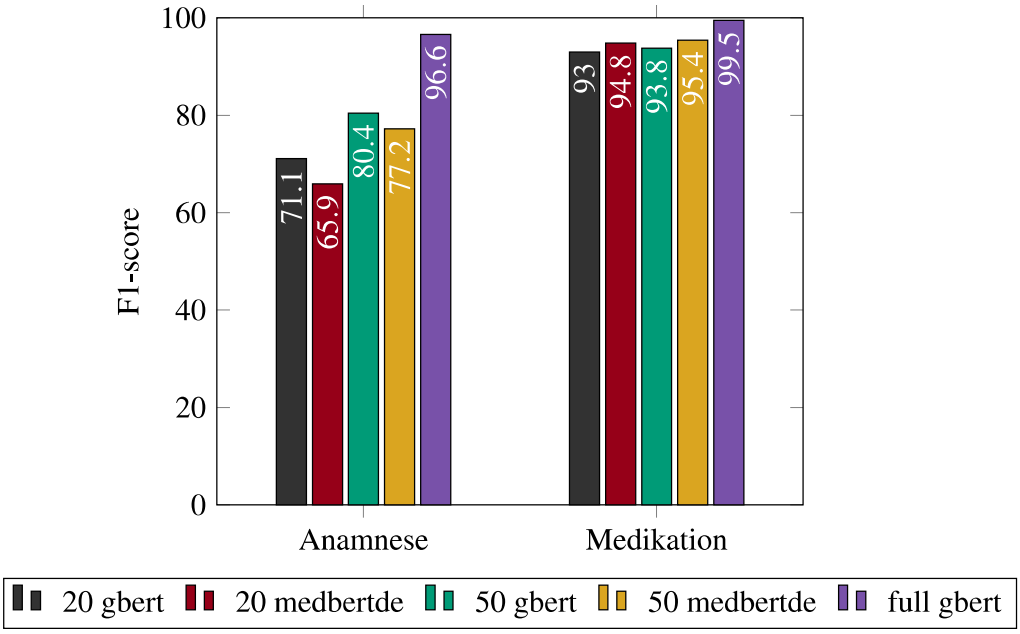


Figure S14: Comparing *gbert-large-comb context* and *medbertde-base context* trained with all templates. F1-score per primary label.

[SEP] über Ihre Patientin Frau Martina Mustermann geboren am 12.12.1999 wh. 2000 Musterstadt Musterstr. 1 die sich am in unserer Ambulanz vorstellte. [SEP] **Anamnese:**

English: [SEP] regarding your patient Mrs. Martina Mustermann, born on December 12, 1999, residing at Musterstr. 1, 2000 Musterstadt, who presented herself at our outpatient clinic. [SEP] Patient Medical History:

[SEP] über Ihre Patientin Herr Max Mustermann geboren am 12.12.1999 wh. 2000 Musterstadt Musterstr. 1 die sich am in unserer Ambulanz vorstellte. [SEP] **Medikation:**

English: [SEP] regarding your patient Mrs. Martina Mustermann, born on December 12, 1999, residing at Musterstr. 1, 2000 Musterstadt, who presented herself at our outpatient clinic. [SEP] Medication:

Figure S15: Two artificial training samples including English translation with atypical co-occuring context paragraphs. In the first sample, the section title *Anamnese* follows immediately after a *Anrede* sample. In the second example, *Medikation* follows after a *Anrede* sample. Usually *Anrede* is followed by *Diagnose*, rarely by *Anamnese* and never in our data set by *Medikation*.

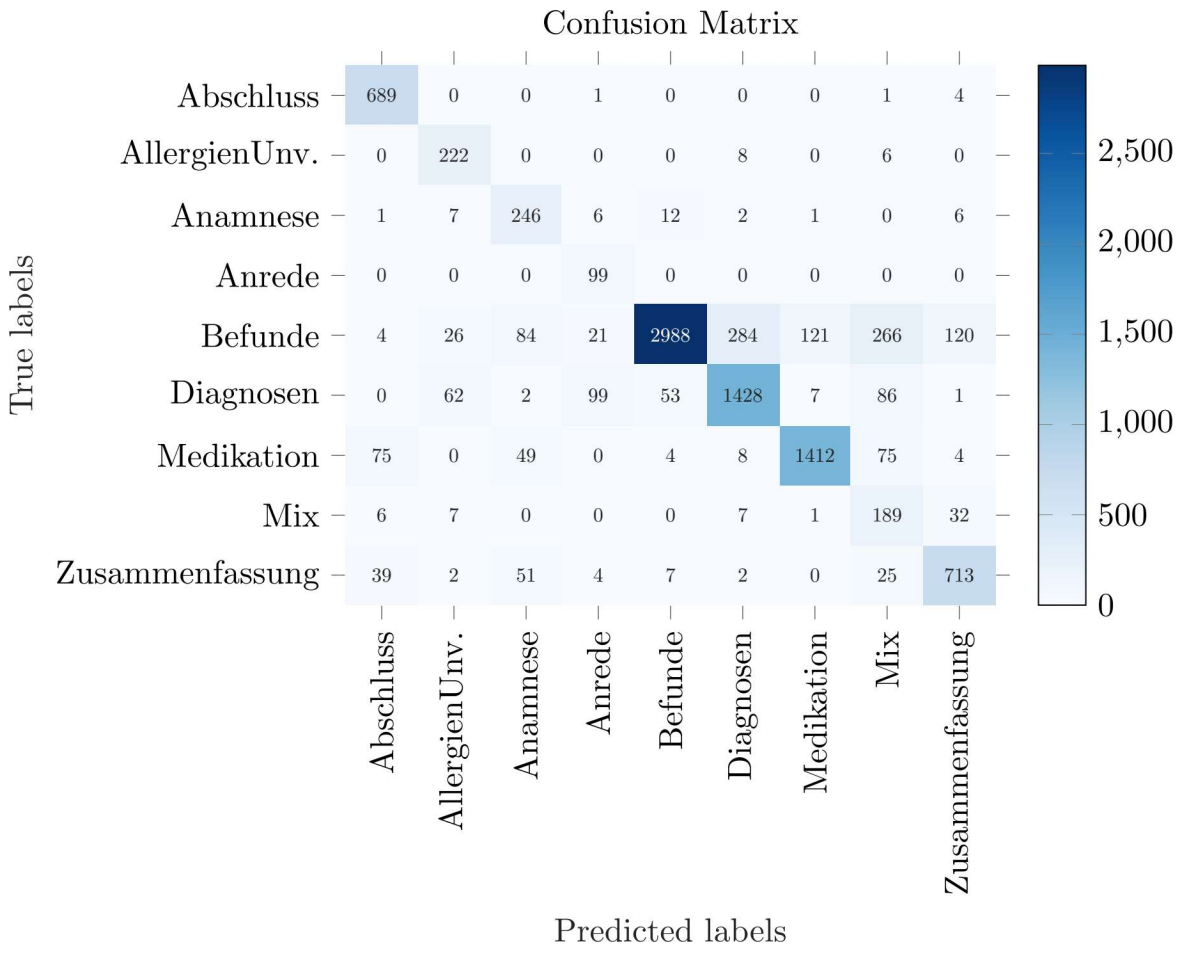


Figure S16: Confusion matrix for *gbert-base-comb context* trained on 20 shots on training set 3 with initial seed 123.

Anamnese:, Diagnosen:, Zusammenfassung:, Körperlicher Untersuchungsbefund:, Aktuell:, Labor:, Ruhe-EKG:, Procedere:, Medikation bei Aufnahme:, Therapieempfehlung:, Beurteilung:, Therapieempfehlung -LRB- von kardiologischer Seite -RRB- :, Befund und Beurteilung:, Transthorakale Echokardiographie:, Maßnahmen:, Befund:, Aktuelle Medikation:, Beurteilender Abschnitt:, Beschreibender Abschnitt:, Zusammenfassende Beurteilung:, Belastungs-EKG:, Kultureller Befund:, Medikation bei Entlassung:, Lokalbefund:, Körperliche Untersuchung:, Allergien:, Ruhe-EKG bei Aufnahme:, Oral:, Kardiovaskuläre Risikofaktoren:, Nächster Termin/Kontrolle:, Procedere/Termine:, Aktuelle Therapie:, Diagnose:, Echokardiographie:, Bisherige Medikation:, Farbduplexsonographie der Gefäße der rechten Leiste:, Kapilläre Blutgasanalyse:, Sonstige Diagnosen:, Therapieempfehlung von kardiologischer Seite:, Nächster Termin/Prozedere:, Befund/Zusammenfassung:, Kommentar:, Indikation für stationären Herzkatheter:, EKG:, Spirometrie:, Wichtig:, Medikation:, Langzeit-EKG vom B-DATE:, Aktuelle/Bisherige Medikation:, Befund / Zusammenfassung:

Figure S17: List of most common section titles: We generated this list by filtering the data set by short sequences including a single ":" at the end.

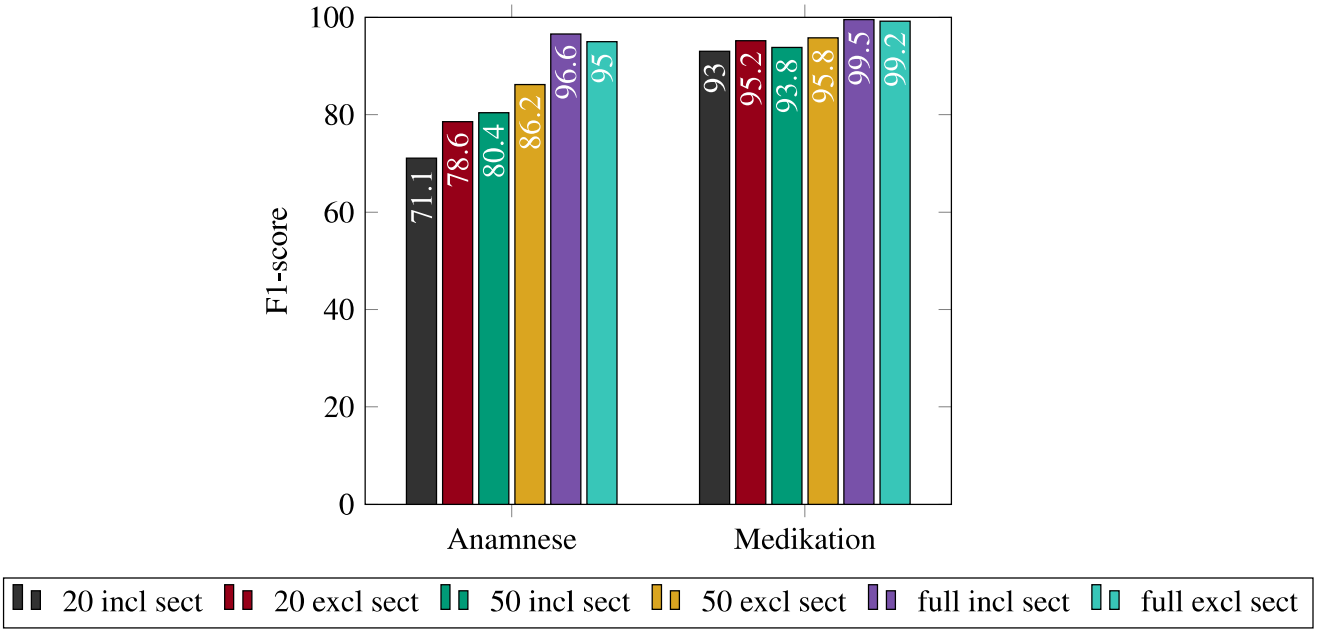


Figure S18: Comparing accuracy scores for *gbert-large-comb context* including and excluding section titles. For reference we show results for SC model trained on full training sets for both scenarios.

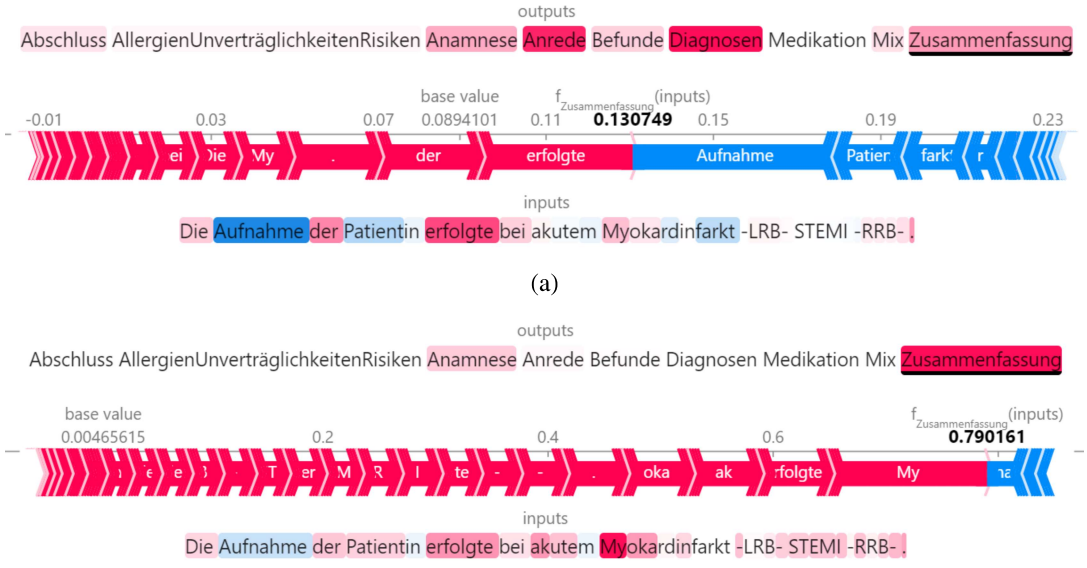


Figure S19: Comparing Shapley values for (a) *gbert-base-comb* context model vs. (b) *gbert-large-comb* context model using a sample without context. Shapley values with respect to predicted label (underlined). Shapley values per sub tokens. Legend: Red: positive contribution, Blue: negative contribution.