DECSELFMASK: Leveraging Unlabeled Text via Self-Relevance-Guided Masking for Decoder-Only Classification

Pietro Ferrazzi[1,2], **Matteo Merler**[1],
Giovanni Bonetta [1], **Alberto Lavelli** [1], **Bernardo Magnini**[1]
[1]Fondazione Bruno Kessler, Trento, Italy
[2]University of Padova, Italy
Correspondence: pferrazzi [at] fbk [dot] eu

Abstract

Classification tasks require annotated data, which can often be expensive, time-consuming, or even unfeasible to collect. This is the case of the medical domain, where large datasets often have few annotated examples. To address this, we propose DecSelfMask (Decoder Self-learning by Masking), an approach to enhance decoder-only performance on classification tasks. We build on common self-learning approaches by leveraging a model to create training examples from unlabeled data to propose a novel relevance-guided masking strategy. We use relevance attribution methods to determine what portions of unannotated texts are relevant for a task. We then create self-supervised training examples by masking out those portions, training the model to reconstruct them via next-token-prediction. We hypothesize that those examples convey knowledge about the structure and semantics of unannotated data that can be useful for downstream performance. We test our approach on 136 tasks from a collection of 1.9M clinical notes from an Italian hospital. We quantify DecSelfMask's impact on downstream tasks on 5 models of different scales and families, including a probing analysis. Experiments show consistent gains, outperforming standard supervised fine-tuning approaches (+19.9 points in Macro F1), synthetic label generation (+12.5), and continual pretraining (+6.3), as well as common baselines.

1 Introduction

Training automatic systems to perform Natural Language Processing tasks requires large amounts of high-quality annotated data (van Engelen and Hoos, 2020; Yakimovich et al., 2021). Annotations collection can be performed by non-experts (Snow et al., 2008), but there are cases where high degrees of expertise are required (Yang et al., 2019). Such processes can be expensive and time-consuming, and in domains like medicine they can represent a major issue (Yakimovich et al., 2021).

Because of these challenges, medical applications often involve an abundance of raw data, and limited annotated examples. Common approaches to tackle these scenarios rely on *self-training* (Wang et al., 2023), where manual annotations are first leveraged to generate synthetic ones, which are then combined with the former to train downstream models (Shi et al., 2023; Luo et al., 2025). Other approaches use raw data to perform *continual pretraining*, aiming to adapt models to the target domain (Singhal et al., 2023; Wu et al., 2024), building on the relation existing between the raw text and the downstream task.

While *self-learning* tries to project such knowledge into unannotated text and expand the task-oriented training set, *continual-pretraining* attempts to gain domain-oriented knowledge from the data itself, without explicitly considering the target tasks. An interesting direction remains underexplored at their intersection, combining the acquisition of knowledge from raw data and the guidance that can be given by annotated examples.

For instance, a target task on *"asthma"* classification can provide guidance to automatically identify portions of a raw text such as *"frequent stitches when breathing"* and *"critical respiratory rate"*, to learn the implicit relation existing between them, and potentially increase downstream performance on the target task itself (Figure 1).

Following this hypothesis, we propose **Decoder Self-learning by Masking** (DECSELFMASK), a method to create training examples from the knowledge embedded in unannotated data. At its core, it is a self-learning-based technique to automatically generate masked examples to train decoder-only models, specifically targeting relevant portions of the input.

Our contributions can be summarized as follows:

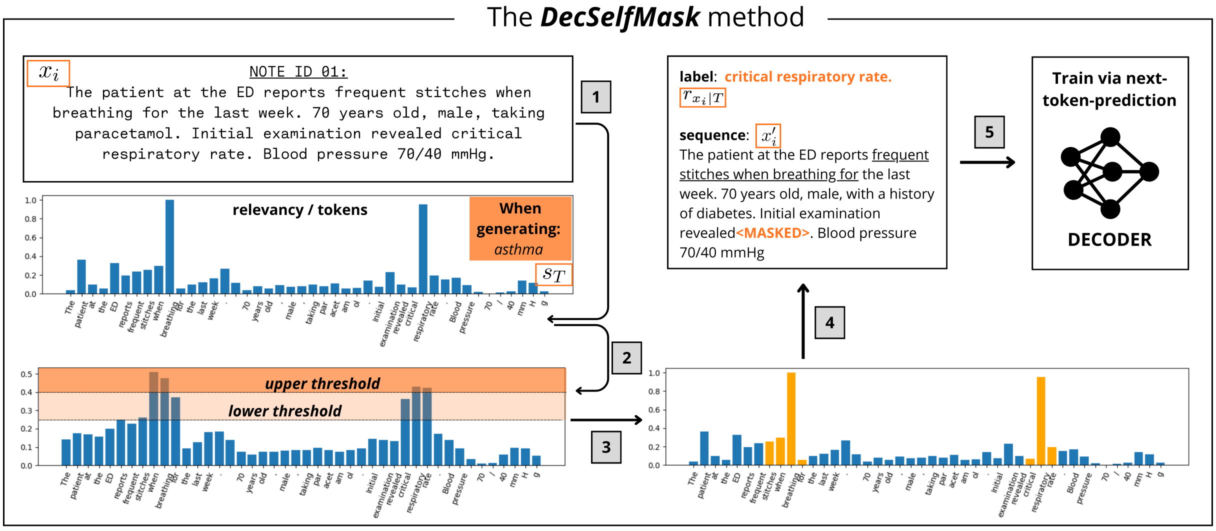

Figure 1: **DECSELFMASK method overview**, exemplified by the task of determining if a patient has asthma. **1) Relevancy calculation**: the raw text x_i is pre-pended to the target text portion s_T (e.g., *asthma*) and given as input to the model. A relevance score is computed for each input token with respect to s_T. **2) Gaussian smoothing**: the relevance scores are smoothed and two thresholds are identified. **3) Groups generation**: portions of relevant text $r_{x_i|T}$ are identified by all consecutive tokens with relevance higher than a lower threshold, where at least one token has relevance higher than the lower threshold. **4) Masking**: training sequences x'_i are generated by masking out relevant portions of text x_i. **5) Training**: decoders are trained via next-token-prediction on the generated examples.

- We propose DECSELFMASK, a method to combine training-based domain adaptation (continual pretraining) with self-supervised learning, aiming to increase performance on classification tasks.
- We evaluate DECSELFMASK on 136 tasks in the medical domain, finding an increase from standard approaches of +19.9 Macro F1 points on average among models scales and families.
- We compare DECSELFMASK to several baselines and state-of-the-art approaches, with gains of +6.3 over continual pretraining and +12.5 over self-learning.

We release the codebase to utilize our framework and replicate the experiments at [ANONIMIZED][1].

2 Related Work

Semi-supervised Learning and Self-training. Automatic systems can be *supervised* or *unsupervised* (Bishop, 2006). When labeled data is present but scarce, mixed *semi-supervised* approaches have shown to be effective (Zhu, 2008). With the rise of LLMs, these methods have shifted toward *self-learning* (Singhal et al., 2023), where models are trained on the labeled examples aiming to generate synthetic-labels from unannotated data, to be then utilized for further training (Gera et al., 2022; Shi et al., 2023; Wang et al., 2023; Luo et al., 2025). While these methods have shown positive impact, they do not explicitly target domain-adaptation dimensions that might be relevant for domain-specific settings.

Medical Domain Adaptation via Continual Pretraining. With the advent of LLMs, there has been a growing interest in models adaptation to the medical domain. The main technique is continual pretraining (CPT) (Ou et al., 2025), where models are exposed to large corpora of raw medical data via next-token-prediction, as this requires no expert labels. Examples are BioGPT (Luo et al., 2022), MedPALM-1 (Singhal et al., 2023), PMC-LLaMA (Wu et al., 2024), MedPalm-2 (Singhal et al., 2025), MedPhi (Corbeil et al., 2025), MedGemma (Sellergren et al., 2026). These methods have shown to be effective by aligning models' knowledge to the target domain, without directly modeling downstream tasks. Notably, others have shown CPT having low-to-no impact on downstream medical tasks (Ferrazzi et al., 2026a) suggesting that alternative approaches should be

[1]For the revision process, our codebase is updated as a *zip* file

explored.

Decoder-only Masking Strategies. Dukić and Šnajder (2024); Huang et al. (2025) propose to act on decoders' causal mask to improve performance on sequence labeling and question answering tasks. More similar to us, Khosla et al. (2025) present a method to enable decoders to generate sentence-level representations, enabling them to perform text in-filling using bi-directional context. By contrast, we do not modify the models' structure in any way.

Relevance Attribution in Transformers. Several methods have been proposed to determine the relevance of the input for a given output generated by a transformer-based model (Ferrando et al., 2024). While perturbation approaches (Covert et al., 2021) and contrastive attribution (Yin and Neubig, 2022) require extensive comparisons and alternative search, gradient-based methods can be applied independently from the context. Hence, we consider AttnLRP (Achtibat et al., 2024) for input relevance attribution in the context of our method.

3 Problem Formulation and Hypothesis

In this work, we propose a method to leverage the structure of unlabeled, domain-specific corpora to solve downstream classification tasks. Our approach relies on the hypothesis that if relevant information can be automatically identified, it can be exploited to construct self-supervised learning objectives, aligned with the downstream task.

Learning from Unlabeled Data. Let us define $X = \{x_i\}_{i=1}^N$ denoting a collection of sequences sampled from a domain-specific corpus, and let $C = \{c_i \in \{c_1, ..., c_K\}\}_{i=1}^N$ denote the relative class labels for the classification task T, which are unknown. Following van Engelen and Hoos (2020), we investigate whether the marginal distribution over the input space $\mathrm{P}(x)$ contains useful information for learning the posterior distribution $\mathrm{P}(c \mid x)$. In our setting, a decoder-only language model $\mathrm{LLM}(c \mid x)$ is used to estimate this conditional distribution. Our central hypothesis, similar to the one behind continual pretraining, is that the conditional distribution $\mathrm{P}(c \mid x)$ is not independent from the structure of the input distribution $\mathrm{P}(x)$. Consequently, we hypothesize that learning regularities, dependencies, and semantic structures within the unlabeled examples sampled from $\mathrm{P}(x)$ can improve performance on the downstream task $T \sim \mathrm{P}(c \mid x)$. However, the relationship between $\mathrm{P}(x)$ and $\mathrm{P}(c \mid x)$ is generally implicit and task-dependent (Oliver et al., 2018). In what follows, we describe our approach to verify this hypothesis.

Linking $\mathrm{P}(\mathbf{c}|\mathbf{x})$ and $\mathrm{P}(\mathbf{x})$ through LLM. Classification tasks rely on the fact that given an input sequence x_i and a task T, there are portions of x_i that are relevant to predict the task-related label c_i. Let

$$r_{x_i|T}(\text{start}, \text{end}) = \{x_{i,j}\}_{j=\text{start}}^{\text{end}} \subset x_i \quad (1)$$

denote a subset of tokens in x_i identified as *relevant*[2] for the downstream task T. Instead of directly learning $\mathrm{LLM}(c_i \mid x_i)$, we define an auxiliary self-supervised objective:

$$\mathrm{LLM}(r_{x_i|T} \mid x_i'), \quad (2)$$

where the model predicts a relevant portion of the input $r_{x_i|T}$ from $x_i' = x_i \setminus r_{x_i|T}$, the input text itself where the relevant portion has been masked out (Figure 1, top right). This formulation resembles self-learning, where the model itself identifies the labels to learn from. Our hypothesis is then framed as follows: the LLM can be trained on the objective defined in Equation 2 (which solely depends on the input $\mathrm{P(x)}$) to better approximate $\mathrm{P(c|x)}$.

In other words, we hypothesize that learning to auto-regressively predict task-relevant portions $r_{x_i|T}$ from x_i' encourages the model to acquire domain-specific knowledge, linguistic structures, and latent correlations that are also beneficial for the downstream task T.

The "Two Peaks" Intuition. We aim to train models on those inputs x_i that convey some relevant knowledge $r_{x_i|T}$ about the downstream task T, according to the structure of Equation 2. On the other hand, we rely on utilizing $r_{x_i|T}$ as a label that needs to be predicted, masking it out from the training sequence x_i'. Then, there is no guarantee that, after masking, x_i' will convey enough task-relevant information to properly predict $r_{x_i|T}$ in the context of T. Therefore, we include in our training set only sequences with two or more disjointed relevant portions $r_{x_i|T}(s_1, e_1)$ and $r_{x_i|T}(s_2, e_2)$ (Figure 1, bottom right). By doing so, we leverage the self-learning framework to its full extent, by training models to predict one portion of text that the model itself identified as relevant to the task, ensuring at least another one exists to provide enough context

[2] For simplicity, in the following we refer to $r_{x_i|T}(\text{start}, \text{end})$ as $r_{x_i|T}$

for this to be possible. We refer to this idea as the "two peaks" framework, given that we only select sequences with at least two peaks of relevance.

Relevance Estimation. Our method relies on the ability to extract portions $r_{x|T}$ that are somehow *relevant* for the downstream task T, to then expose models to that information. Following the concept of *self-learning*, we compute such relevance by leveraging the model itself. We append to each input x_i a short string s_T that describes the task T (in the example of Figure 1, $s_T =$ "$asthma$"). Then, we use AttnLRP (Achtibat et al., 2024) to quantify the relevance attributed by the model itself to each token in x_i if it was generating s_T. For a given sequence of input tokens $x_i = \{x_{i,1}, ...x_{i,n}\}$, AttnLRP provides a relevance score for each token $\text{rel}_{x_i|s_T} = \{\text{rel}_{x_{i,1}|s_T}, \ldots, \text{rel}_{x_{i,n}|s_T}\}$.

Training Data Selection. We define a sequence x_i as relevant for a task T if there exist at least two disjoint subsequences where all tokens have relevance higher than a lower threshold t_l, and at least one token has relevance higher than an upper threshold t_u (following the "two peaks" framework). Formally, each of the two disjoint subsequences $r_{x_i|T}(s, e)$ complies with the following conditions:

$$
\begin{aligned}
&\text{cond}_1(s, e) : G(\text{rel}_{x_{i,j}|s_T}) \geq t_l \quad \forall j \in [s, e] \\
&\text{cond}_2(s, e) : \exists j \in [s, e] \text{ s.t. } G(\text{rel}_{x_{i,j}|s_T}) \geq t_u,
\end{aligned} \tag{3}
$$

where $G()$ is a gaussian smoothing function on the relevance scores (Figure 1, step 2).

Since $\text{rel}_{x_{i,j}|s_T}$ is defined over tokens, $r_{x_i|T}$ spans may not align with full words and can split sub-word tokens. For smoothing purposes, we extend each span to cover the complete words intersecting the selected token range.

The objective proposed in Equation 2, combined with the training data defined in this manner, acts as a form of task-oriented self-supervision: the model extracts supervision signals directly from unlabeled data by leveraging the fact that it has identified as relevant for the task at least another portion of text.

4 Experimental Setup

To verify the hypothesis defined in Section 3, we require a large dataset of **unlabeled data** together with a smaller set of **expert-annotated data** for medical classification tasks. We leverage the SGB dataset of medical text presented by Ferrazzi et al. (2026b), released with a permissive license. This comprises around 1.9 million anonymized clinical notes coming from the Emergency Department of the San Giovanni Bosco Hospital in Italy, including all notes produced over a three-year span.

Ferrazzi et al. (2026c) further proposed the Case Report Form (CRF) filling task built on top of the SGB dataset. The task consists of filling a list of 136 medical items with the correct values, given a clinical note as input. We reshape the task to make it resemble pure **classification** and obtain a set of 136 individual classification tasks, one for each of the original CRF items. Each classification task is characterized by three or more possible labels. For instance, the classification task defined by the string s_T "*heart rate*" consists in determining whether a patient is "*bradycardic*", "*normocardic*", "*tachycardic*", or if there is not enough information to determine so ("*unknown*"). We obtain the dataset extended to 6404 notes[3], each annotated for one or more relevant tasks, resulting in $61k$ pairs of note-annotation in total. Each task has an average of 440 annotated examples per task (variance of 147). A comprehensive list of all tasks and their label space is reported in Table 5 (Appendix).

Furthermore, we define two **held-out classification tasks** on notes from the SGB dataset, determining if a patient presents any "*chronic*" condition, or determining whether a patient is entering the emergency department with "*dyspnea*" or "*loss of consciousness (tloc)*". We reserve the two tasks for evaluating whether the knowledge acquired during DECSELFMASK training can transfer to previously unseen tasks. We believe this setting is particularly relevant in real-world medical environments, where new classification needs may emerge over time. All datasets are described in Table 1.

Relevancy Calculation. We follow the procedure described in Section 3 to generate the masked training data (Figure 1). For each task T (out of the 136), we need to define the target s_T to base the relevance generation on. In the datasets we use, each T comes with a short text (one to five words) that defines the dimension to classify the note on (e.g. *heart rate*, Table 5 for full list). Therefore, we can use it as s_T, making sure the relevance calculation is based on text that is relevant for T itself. For each task, we append s_T to each clinical note in the SGB dataset (template is provided in Appendix C).

The calculation of the relevance via AttnLRP

[3] Can be obtained via email from the CRF Shared Task org.

Dataset	Description	Use	n	words	n tasks
SGB	Anonymized raw clinical notes from the emergency department of an Italian hospital.	Generation of the DECSELFMASK training examples.	1.9M	125.7M	—
CRF	Annotated clinical notes from the SGB dataset on 136 medical items. Each note is annotated once for each relevant item.	Guidance on how to construct the DECSELFMASK training examples; main evaluation.	61k	4.7M	136
Chronicity	Annotated clinical notes from the SGB dataset determining whether a patient has chronic conditions.	Evaluate DECSELFMASK on tasks not involved in its definition and training.	2587	335k	1
T-D	Annotated clinical notes from the SGB dataset determining whether a patient is admitted to the emergency department with loss of consciousness or dyspnea.	Evaluate DECSELFMASK on tasks not involved in its definition and training.	1713	331k	1

Table 1: **Datasets used in our experiments**, with number of examples *n*, *words*, and *tasks* (where applicable). *SGB* is used for DECSELFMASK learning, to improve performance on classifying notes on the 136 *CRF* items. *Chronicity* and *T-D* are not directly involved in the DECSELFMASK process, but left for further testing.

must be performed over one single generated token; however, all 136 s_T texts are multi-token sequences. While we could average the relevance calculated for each generated token, this would result in repeated runs over the input, increasing the computational burden. Instead, we use the relevance computed on the middle token of s_T as a computationally efficient proxy. See Appendix A for the validation of this choice. For the same reasons, we do not utilize the whole dataset, and limit to a randomly selected set of half a million clinical notes, and two tasks for each note. Overall, we run relevance calculation on one million prompts. This step takes 30 GPU hours on one NVIDIA L40S.

DECSELFMASK Training Sequences Generation. Once the relevance is obtained for all notes, we perform Gaussian smoothing (kernel size=3, sigma=1), and set an upper (0.4) and lower (0.2) thresholds to identify the groups of relevant tokens. Then, we drop all sequences that do not contain at least "two peaks", and perform de-duplication, as different s_T texts might have identified the same token groups as relevant for different tasks T, which would result in identical training sequences.

For each note containing two relevant spans, we generate training examples by masking one span at a time (Figure 1, top right). Due to the large number of generated examples, we further filter the dataset by retaining only sequences in which another relevant span occurs to the left of the masked region. This has two reasons: (i) two relevant sequences from the same clinical note most likely contain similar information, and (ii) encouraging the model to infer the masked content from preceding contextual information, thus mimicking the autoregressive nature of the downstream task.

We collect all the sequences found to be useful for all 136 tasks in one single dataset of $1.03M$ examples. Each example is composed by a clinical note where a portion has been masked out (to be used as *input*), the text s_T used for the masking process, the text in the masked portion (to be used as *label*), and the valid labels.

DECSELFMASK Training. To verify if the generated training pairs of input and output can be utilized as sources of knowledge about the downstream classification tasks they were designed for, we analyze three model families in their instructed versions: `Llama3`, `Qwen3` in their ~`1B` and `8B` versions, and `Gemma3-4B` to represent a mid size between the previous ones. We train all parameters for one epoch using the hyperparameters described in Appendix B on two NVIDIA H200 GPUs (requiring a total of 68 GPU hours). The learning objective is next token prediction over the prompt structure presented in Appendix C. By doing so, we obtain 5 DECSELFMASK-models, which can be used to analyze the effect of our method.

5 Evaluation and Discussion

Downstream Tasks Training. We compare the base models with their DECSELFMASK versions, to determine the impact of DECSELFMASK training. We aim to adapt both versions to the downstream tasks, and determine which performs better. To do so, we perform supervised fine-tuning (SFT) on the 136 downstream tasks. We divide the CRF dataset into an 85-15 train-test split. For both set-

Model	SFT			Probing			Δ_{best}
	Base	**DSM**	Δ_{DSM}	**Base**	**DSM**	Δ_{DSM}	
Llama3-1B	21.9 ± 1.0	19.69± 1.0	-2.2	34.1± 1.1	47.0 ± 1.3	+12.9	+25.2
Qwen3-1.7B	30.9 ± 1.2	22.3 ± 1.0	-8.6	33.3 ± 1.2	46.8 ± 1.3	+13.5	+15.9
Gemma3-4B	37.5 ± 1.2	44.4± 1.3	+7.0	36.1 ± 1.2	47.5 ± 1.3	+11.4	+10.0
Llama3-8B	35.3 ± 1.0	42.9 ± 1.2	+7.6	55.6 ± 1.3	55.8 ± 1.3	+0.2	+20.5
Qwen3-8B	32.5 ± 1.2	40.4 ± 1.2	+7.9	52.5 ± 1.3	60.3 ± 1.2	+7.7	+27.8

Table 2: **Impact of DecSelfMask training across task-adaptation strategies.** The *SFT* columns report performance for supervised fine-tuning of the base or DSM models, while the *Probing* columns report performance when using a classification head on the final layer. For each strategy, Δ_{DSM} is the difference between the DSM model and its corresponding base model. The final column reports the performance gain of the best DSM adaptation strategy (either SFT or probing) over the standard SFT baseline applied to the base model. Results are reported as Macro F1 with 95% confidence intervals.

tings, we train one single multi-task model for all tasks at once, using LoRA (Hu et al., 2022), for one epoch, with the hyperparameters described in Appendix B, for a total of 30 GPU hours.

We calculate performance on all 136 tasks. As the aggregation metric, we use the macro F1 following previous work (Ferrazzi et al., 2026c), to summarize the results on all tasks in one single metric. We compare the performance of base models which undergo task-oriented SFT to our DecSelfMask-models that undergo the same fine-tuning.

As shown in Table 2 (SFT columns), we observe that DecSelfMask was detrimental for small models ($\sim$1B), with an average decrease of performance of -5.4. On the other hand, it resulted in a homogeneous increase for bigger models, with a $+7.5$ positive impact.

We hypothesize that this behavior is not due to small models failing to acquire useful knowledge with DecSelfMask, but rather to their limited capacity to coherently generate tokens that reflect this knowledge when prompted to generate the output. In other words, smaller models may successfully encode information that is beneficial for the downstream task, yet lack the representational capacity required to effectively transfer or exploit this knowledge after autoregressive SFT. To verify this hypothesis, further analysis on `DecSelfMask-Llama3-1B` and `DecSelfMask-Qwen3-1.7B` is required.

Probing. To better understand why training with DecSelfMask improves downstream performance only for bigger models, we perform a probing analysis by training a shallow classifier on top of the last hidden layer of the small models ($\sim$ 1B). We add a classification head on the last hidden representation, composed by a linear layer (output size of 256), a ReLU activation with 0.1 dropout, and an output layer to map the 256 neurons to the output classes. We train this head with the hyperparameters described in Appendix B, for an overall of 24 L40S GPU hours. The results are reported in Table 2 (last four columns). We observe a consistent positive effect, suggesting that DecSelfMask successfully injects useful information into the learned representations, even in cases where this does not translate into improved downstream autoregressive generation performance. The models that underwent our training result in a homogeneous average increase of $+13.2$ points.

From SFT to Probing. The probing analysis suggests that even smaller models learn an useful representation for the downstream tasks. Furthermore, by comparing the results obtained via SFT and probing, we observed that the potential of the models on these tasks goes much beyond what they exhibit after standard, auto-regressive SFT (an effect that has been analyzed by Lyu et al. (2026)). In fact, `Llama3-1B` with a classification head before SFT achieves a Macro F1 $+11.2$ points higher than after standard SFT and without the head ($+2.4$ for `Qwen3-1.7B`). Building on this evidence, we train a classification head on top of the larger models as well, to verify if they could achieve better performance. We use the same structure and training parameters described above, obtaining the results shown in Table 2.

We highlight two aspects: first, when adapting models to downstream tasks by adding a classification head, DecSelfMask gives an average

Model	Tloc-Dyspnea			Chronicity			Average Gain	
	SFT Base	**CH Base**	**CH DSM**	**SFT Base**	**CH Base**	**CH DSM**	Δ_{CH}	Δ_{SFT}
`Llama3-1B`	50.5	79.5	84.0	65.8	65.8	69.5	+4.1	+18.6
`Qwen3-1.7B`	36.9	78.1	86.7	68.9	65.3	68.3	+5.8	+24.6
`Gemma3-4B`	50.9	85.5	82.6	63.5	66.3	70.1	+0.4	+19.2
`Llama3-8B`	71.6	80.2	81.3	64.9	64.5	68.1	+2.3	+6.5
`Qwen3-8B`	47.2	79.3	80.3	56.9	64.7	68.8	+2.6	+22.5

Table 3: **Transfer to Held Out Classification Tasks.** We evaluate whether representations learned through DECSELFMASK training transfer to classification tasks not used when constructing the DECSELFMASK training sequences. For each task, we report F1 scores for supervised fine-tuning of the base model (*SFT Base*), a classification head trained on the base model (*CH Base*), and a classification head trained on the corresponding DECSELFMASK model (*CH DSM*). Δ_{CH} is the average F1 gain of CH DSM over CH Base across the two tasks. Δ_{SFT} is the average F1 gain of CH DSM over SFT Base across the two tasks.

improvement of $+9.1$ in Macro F1, showcasing that the training procedure effectively injects useful knowledge into the models' representations from the unannotated data. Second, the average performance gain moving from auto-regressive SFT on the base model to classification heads on the DECSELFMASK version is $+19.9$ points, showing this approach to be the most effective.

Held Out Tasks. We reserve two tasks to evaluate whether the knowledge acquired during DECSELFMASK training on the 136 medical classification tasks can transfer to new classification settings. We employ two datasets *Chronicity* and *T-D* described in Section 4 and in Table 1. The evaluation metric for both tasks is the F1 score.

We build on the finding that classification heads overperform SFT, and follow this to adapt base models and their DECSELFMASK counterparts. In addition, we provide for comparison standard SFT results on the base models, and report all findings in Table 3. The probing strategy showcases that DECSELFMASK results in higher F1 ($+3.0$ on average across all models) when compared to the base, and it presents an even higher increase of $+18.3$ when compared with the base with SFT.

Baselines. We compare DECSELFMASK to other methods, to contextualize its efficacy. The natural baseline for text classification tasks are **encoder-only models**. We fine-tune on the 136 CRF downstream tasks `BERT` and its multilingual version `multilingual-BERT` (Devlin et al., 2019), `BioClinicalBERT` (Alsentzer et al., 2019), `ModernBERT` (Warner et al., 2025), and a version of BERT explicitly trained for Italian (Schweter, 2020). The models are trained to receive as input a clinical note and the text describing the task s_T (e.g., "*heart rate*") with the hyperparameters defined in Appendix B.

Moreover, we provide the performance of the decoder model that showed the best results in our experiments (`Qwen3-8B`), prompted **0- and 2-shot**. For 0-shot, we run inference 3 times with slightly different system prompts in order to account for instabilities, following Sclar et al. (2024), and report the average. For 2-shot,we randomly sample the examples, and run inference 3 times to mitigate bias due to their choice. Results are presented in Table 4, showing preference for encoders (up to $+13.4$ points), which are still outperformed by DECSELFMASK.

State of the Art Comparison. Our approach relies on training on the unannotated, raw data. As described in Section 2, the typical approach to expose models to large amounts of raw data is **continual pretraining** (CPT), whose underlying hypothesis inspires DECSELFMASK. We then perform CPT on a model on the same (unmasked) raw clinical notes included in our method. We train using the same hyperparameters as in DECSELFMASK, with the standard next-token-prediction objective. While CPT has a positive impact, it still falls short w.r.t. DECSELFMASK (-6.3 points).

Finally, we include a standard **self-learning** approach, which we take inspiration from to design the training objective of Equation 2. We train a model on the downstream tasks via SFT on the annotated CRF data (similar to the SFT Base models from Tables 2 and 3), and then use it to generate synthetic labels for the unannotated SGB dataset. Then, we train the original base model on the combination of the original and the synthetic labeled

Model	Macro F1
ClinicalBERT	34.1 ± 1.2
multilingual-BERT	34.2 ± 1.2
BERT	36.8 ± 1.2
BERT-Italian	38.2 ± 1.3
ModernBERT	44.1 ± 1.3
Qwen3-8B + 0-shot	27.7 ± 0.6
Qwen3-8B + 2-shot	30.7 ± 0.7
Qwen3-8B + SFT	32.5 ± 1.1
Qwen3-8B + CH	32.5 ± 1.1
Qwen3-8B + CPT + SFT	33.5 ± 1.3
Qwen3-8B + CPT + CH	54.0 ± 1.4
Qwen3-8B + synthetic SFT	47.8 ± 1.3
Qwen3-8B + random mask + SFT	30.4 ± 1.2
Qwen3-8B + random mask + CH	56.4 ± 1.1
DECSELFMASK-Qwen3-8B-CH	**60.3** ± 1.2

Table 4: **Comparison to baselines and state-ot-the-art** for the 136 CRF tasks compared with our best performing model (DECSELFMASK-Qwen3-8B-CH with a classification head CH). We provide results for 0- and 2-shot; continual pretraining (*CPT*); self-learning where the model is trained via SFT on synthetic labels generated by itself (*synthetic SFT*). We also showcase the performance of applying DECSELFMASK with *random mask*ing instead of relying on relevance attribution.

data via SFT. While outperforming the base model adapted to the task via SFT (+15.3 points), it still falls behind our approach (−12.5).

Overall, these results highlight that our approach obtains better performance than previous methods, by effectively integrating knowledge embedded in raw data and shaping it through a procedure that enhances classification performance.

DECSELFMASK further analysis. Our training procedure modifies the model's parameters to encode information extracted from unlabeled data, effectively steering the model's internal representations toward task-relevant patterns present in the raw corpus. To better understand the effect of this training procedure, we analyze how it alters the relevance assigned to different portions of the input. More specifically, we provide the model with the same prompt used during the relevance extraction phase of our pipeline, and compute the relevance attributed to the input tokens while generating the target item using the same AttnLRP score used when constructing the dataset.

This analysis reveals notable qualitative differences between the base model and the corresponding DECSELFMASK variant. In several cases, our models assign higher relevance to portions of the text that are semantically more aligned with the target concept (e.g, Figures 3 -7 in Appendix).

We quantify this effect by leveraging the labeled clinical notes, where annotators also identified the portions of text considered relevant for each of the 136 classification tasks. We measure the change in relevance assigned to these ground-truth spans when moving from the base model to the corresponding DECSELFMASK variant. For Qwen-8B, we observe an average relevance increase of 37.2% over the annotated spans, suggesting that the proposed self-supervised training objective effectively steers the model towards assigning higher importance to actually relevant portions of the input.

In addition, we provide results obtained by ablating the relevance attribution step, where we train the model to autoregressively reconstruct sequences randomly masked. Table 4 (last three rows) highlights that the step has a significant impact.

6 Conclusion

In this work, we introduce DECSELFMASK, a task-oriented self-supervised framework to leverage unlabeled, domain-specific corpora to improve the downstream classification performance of decoder-only models, combining the assumptions underlying *self-learning* and *continual pretraining*.

First, we propose to use input attribution methods to mask out task-relevant portions of unlabeled text, and train to autoregressively reconstruct them, encouraging models to learn domain-specific patterns from raw data, guided by relations identified as relevant to the target tasks by models themselves.

Second, we evaluate DECSELFMASK on 136 tasks derived from 1.9M clinical notes. We adapt models to the tasks with a classification head, as probing experiments show consistent improvement over standard SFT across model families and scales (+19.9 F1 Macro points), also showing similar effectiveness on two held-out tasks ad well.

Third, we compare DECSELFMASK to standard *continual pretraining* (+6.3) and *self-learning* approaches (+12.5), where results further highlight the effectiveness of our approach.

Overall, our findings suggest that DECSELFMASK is an effective way to transform unlabeled domain-specific corpora into useful training signals for decoder-only language models.

Limitations

Our experiments are limited to Italian clinical notes and relatively small decoder-only models, although results suggest that larger models may benefit more from DECSELFMASK. Due to the computational cost of AttnLRP, we train on a subset of the original 2M-note corpus and perform joint self-supervised training across all tasks rather than task-specific training. Additionally, the proposed framework assumes that downstream tasks can be represented through short textual descriptions s_T. Future work should explore more efficient relevance extraction methods, alternative masking objectives, and extensions to broader task settings and larger models.

Ethical considerations

Although our approach aims to reduce the dependency on expensive manual annotation, models trained on medical data may still inherit biases, incompleteness, or inaccuracies present in the underlying clinical records. Consequently, the proposed models should not be considered a substitute for professional medical judgment.

Additionally, relevance attribution methods may produce imperfect or misleading explanations of model behavior. Therefore, the relevance analyses presented in this work should be interpreted as approximations of model reasoning rather than faithful causal explanations.

References

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. 2024. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 135–168. PMLR.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer, Berlin.

Jean-Philippe Corbeil, Amin Dada, Jean-Michel Attendu, Asma Ben Abacha, Alessandro Sordoni, Lucas Caccia, Francois Beaulieu, Thomas Lin, Jens Kleesiek, and Paul Vozila. 2025. A modular approach for clinical SLMs driven by synthetic data with pre-instruction tuning, model merging, and clinical-tasks alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19352–19374, Vienna, Austria. Association for Computational Linguistics.

Ian Covert, Scott Lundberg, and Su-In Lee. 2021. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David Dukić and Jan Šnajder. 2024. Looking right is sometimes right: Investigating the capabilities of decoder-only LLMs for sequence labeling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14168–14181, Bangkok, Thailand. Association for Computational Linguistics.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. 2024. A primer on the inner workings of transformer-based language models. *Preprint*, arXiv:2405.00208.

Pietro Ferrazzi, Mattia Franzin, Alberto Lavelli, and Bernardo Magnini. 2026a. Small llms for medical nlp: a systematic analysis of few-shot, constraint decoding, fine-tuning and continual pre-training in italian. *Preprint*, arXiv:2602.17475.

Pietro Ferrazzi, Mattia Franzin, Alberto Lavelli, and Bernardo Magnini. 2026b. Small llms for medical nlp: A systematic analysis of few-shot, constraint decoding, fine-tuning and continual pre-training in italian. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 2440–2457, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Pietro Ferrazzi, Soumitra Ghosh, Alberto Lavelli, and Bernardo Magnini. 2026c. Overview of the crf 2026 shared task on clinical case report forms filling. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 245–254. ELRA Language Resources Association (ELRA). 12 May 2026.

Ariel Gera, Alon Halfon, Eyal Shnarch, Yotam Perlitz, Liat Ein-Dor, and Noam Slonim. 2022. Zero-shot text classification with self-training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1119, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shahriar Golchin, Mihai Surdeanu, Nazgol Tavabi, and Ata Kiapour. 2023. Do not mask randomly: Effective domain-adaptive pre-training by masking in-domain keywords. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 13–21, Toronto, Canada. Association for Computational Linguistics.

Yuxian Gu, Zhengyan Zhang, Xiaozhi Wang, Zhiyuan Liu, and Maosong Sun. 2020. Train no evil: Selective masking for task-guided pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6966–6974, Online. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Zixian Huang, Xinwei Huang, Ao Wu, Xiaxia Wang, and Gong Cheng. 2025. Transforming decoder-only models into encoder-only models with improved understanding capabilities. *Knowledge-Based Systems*, 309:112907.

Savya Khosla, Aditi Tiwari, Kushal Kafle, Simon Jenni, Handong Zhao, John Collomosse, and Jing Shi. 2025. MAGNET: Augmenting generative decoders with representation learning and infilling capabilities. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27328–27346, Vienna, Austria. Association for Computational Linguistics.

Junyu Luo, Xiao Luo, Xiusi Chen, Zhiping Xiao, Wei Ju, and Ming Zhang. 2025. Semi-supervised fine-tuning for large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2795–2808, Albuquerque, New Mexico. Association for Computational Linguistics.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Jian Lyu, Jingfeng Xue, Weizhi Meng, Weijie Han, Junbao Chen, and Zeyang Liu. 2026. Systematic design choices for fine-tuning text-classification models: Projection space, task instructions, and label encoding. *Applied Soft Computing*, 187:114341.

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. 2020. The e3c project: Collection and annotation of a multilingual corpus of clinical cases. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 190–196.

Avital Oliver, Augustus Odena, Colin Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. 2018. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Yixin Ou, Yunzhi Yao, Ningyu Zhang, Hui Jin, Jiacheng Sun, Shumin Deng, Zhenguo Li, and Huajun Chen. 2025. How do LLMs acquire new knowledge? a knowledge circuits perspective on continual pretraining. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19889–19913, Vienna, Austria. Association for Computational Linguistics.

Nafis Sadeq, Canwen Xu, and Julian McAuley. 2022. InforMask: Unsupervised informative masking for language model pretraining. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5866–5878, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Stefan Schweter. 2020. Italian bert and electra models.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models'sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, volume 2024, pages 25055–25083.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2026. Medgemma technical report. *Preprint*, arXiv:2507.05201.

Zhengxiang Shi, Francesco Tonolini, Nikolaos Aletras, Emine Yilmaz, Gabriella Kazai, and Yunlong Jiao. 2023. Rethinking semi-supervised learning with language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5614–5634, Toronto, Canada. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Jesper E. van Engelen and Holger H. Hoos. 2020. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, and Y. Wang. 2024. PMC-LLaMA: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pages 21875–21895.

Artur Yakimovich, Anaël Beaugnon, Yi Huang, and Elif Ozkirimli. 2021. Labels in a haystack: Approaches beyond supervised learning in biomedical applications. *Patterns*, 2(12):100383.

Yinfei Yang, Oshin Agarwal, Chris Tar, Byron C. Wallace, and Ani Nenkova. 2019. Predicting annotation difficulty to improve task routing and model performance for biomedical information extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1471–1480, Minneapolis, Minnesota. Association for Computational Linguistics.

Qinyuan Ye, Belinda Z. Li, Sinong Wang, Benjamin Bolte, Hao Ma, Wen-tau Yih, Xiang Ren, and Madian Khabsa. 2021. On the influence of masking policies in intermediate pre-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7190–7202, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kayo Yin and Graham Neubig. 2022. Interpreting language models with contrastive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 184–198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiaojin Zhu. 2008. Semi-supervised learning literature survey. Technical Report 1530, University of Wisconsin–Madison.

A Relevance Computation Details

Encoder-only Masking Strategies. The method we present is inspired by language masking training strategies used for encoder-only models. It has been shown that the choice of what portions of text to mask is important when training encoders using masked language modeling (Gu et al., 2020; Ye et al., 2021; Sadeq et al., 2022; Golchin et al., 2023). We take inspiration from this paradigm, shifting it to autoregressive generative models.

Middle Token Choice. As the cost of calculating the relevance using AttnLRP for all tokens in the output increases linearly with the length of the output, we approximate the average relevance if the input on a certain output using one single output token. A practical challenge arises when the target label is composed of multiple tokens: relevance can be computed using the hidden states associated with different decoding steps, potentially leading to different attribution patterns. In particular, it is unclear whether relevance should be estimated using the first generated token, the last one, or by aggregating information across all label tokens.

To investigate this aspect, we conduct a validation experiment on a medical Named Entity Recognition (NER) dataset (Magnini et al., 2020). For each annotated entity, we construct a prompt of the form:

```
You are an entity extractor.
{text} {label}
```

and evaluate whether the relevance assigned by the model to the tokens in `{text}` corresponding to the target entity is significantly higher than what would be expected under a uniform relevance distribution.

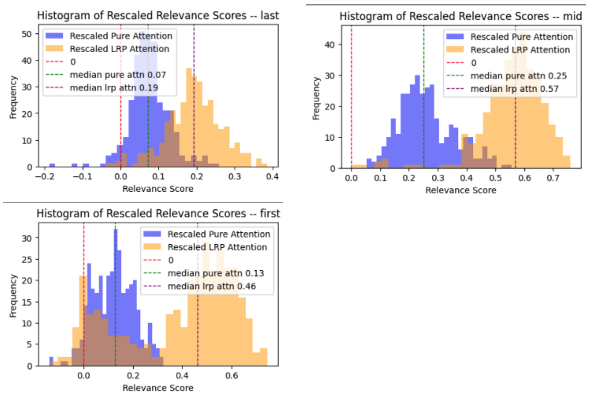

Figure 2: **Comparison between the relevance attributed to ground truth portions of text** i) by *Pure Attention* (i.e., average attention across all layers), and AttnLRP, ii) when utilizing different tokens of s_T for the relevance calculation. The x axis represents the increase in relevance that each method is assigning to the ground truth textual portion with respect with uniform distribution (i.e, the more shifted to the right, the better). The best performing configuration is using the *middle* token with AttnLRP (top-right histogram).

AttnLRP Choice. We compare two relevance estimation methods, the average of the attention scores among all layers, and AttnLRP, by computing token relevance using the hidden states associated with: (i) the first token of the label, (ii) the last token of the label, and (iii) the middle token of the label. Our results show that using the middle token consistently produces the strongest localization signal, yielding higher relevance scores over the entity span (Figure 2). This suggests that intermediate decoding steps provide a more stable representation of the semantic content of multi-token labels compared to boundary tokens.

By doing so, we also validate the usage of AttnLRP, which shows the best performance in terms of relevance attribution to the ground truth in the original text.

B Training parameters

DecSelfMask training We train using `deepspeed Zero2` and `flash attention`, on an effective batch size of 128, using AdamW optimizer with a learning rate of $4 \cdot 10^{-5}$ (cosine scheduler), weight decay of 0.001, warmup ratio 0.1.

Supervised fine-tuning For Supervised fine-tuning, we use a batch size of 64, learning rate of $3 \cdot 10^{-4}$, warmup ratio 0.1, using AdamW optimizer with cosine scheduler and epsilon $1 \cdot 10^{-7}$. We use LoRA (Hu et al., 2022), with r 32, alpha 64, and dropout 0.05. We construct SFT sequences according to the template in Appendix C.

Probing We train for 20 epochs, with learning rate $1 \cdot 10^{-}3$ and the AdamW optimizer.

Encoder-only baseline Training of encoder-only baselines is performed for 10 epochs, with an AdamW optimizer, and a learning rate of $2 \cdot 10^{-5}$. Initial experiments showed that increasing the number of epochs would not provide any benefit.

C Prompts

Relevance calculation We compute the relevance by using the following prompt template:

```
{clinical note}
Given the patient's medical history, the
following is observed {s_T}.
```

Note that models' special tokens are added where required, but they are excluded from the relevance calculation, following the findings of Xiao et al. (2024), as well as the added text.

DecSelfMask training The self-training is performed using he following template:

```
Fill in the masked word in the following
sentence.
{clinical note}
Given the patient's medical history, the
following is observed {label}
```

Note that the simple system prompt is provided to leverage models instruction following, and models' special tokens are added where required.

Supervised fine-tuning When fine-tuning DecSelfMask models on the downstream CRF classification tasks (each task identified by a short description s_T), we combine the data examples according to the following template:

```
Fill in the masked word in the following
sentence.
{masked clinical note}
<crf_item> {s_T}? {masking token} <\crf_item>
<options>{options}<\options>
{label}
```

The masking token is a model-dependent special token that is used to substitute the masked text.

On the other hand, when fine-tuning base models (e.g., `Qwen/Qwen3-8B`) on the same tasks, we initially kept a different template, removing the masking token. Then, we observed that changing the prompt did not impact the performance, and kept the previous versions for all. This finding is consistent with Lyu et al. (2026).

options	**items**
A; V; P; unknown	['level of consciousness']
bradycardic; normocardic; tachycardic; unknown	['heart rate']
bradypneic; eupneic; tachypneic; unknown	['respiratory rate']
certainly active; possibly active; certainly not active; unknown	['active neoplasia']
certainly chronic; possibly chronic; certainly not chronic; unknown	['chronic pulmonary disease', 'chronic respiratory failure', 'chronic cardiac failure', 'chronic renal failure', 'chronic metabolic failure', 'chronic rheumatologic disease', 'chronic dialysis']
current; past; unknown	['presence of respiratory distress']
hypotensive; normotensive; hypertensive; unknown	['blood pressure']
hypothermic; normothermic; hyperthermic; unknown	['body temperature']
measured; unknown	['spo2', 'ph', 'pa02', 'pac02', 'hc03-', 'lactates', 'hemoglobin', 'platelets', 'leukocytes', 'c-reactive protein', 'blood sodium', 'blood potassium', 'blood glucose', 'creatinine', 'transaminases', 'inr', 'troponin', 'bnp or nt-pro-bnp', 'd-dimer', 'blood calcium', 'serum creatinine kinase', 'blood alcohol', 'blood drug dosage', 'urine drug test']
pos; neg; unknown	['carotid sinus massage', 'supine-to-standing systolic blood pressure test', 'blood in the stool', 'sars-cov-2 swab test']
short; long; unknown	["duration of the patient's consciousness recovery", "duration of the patient's unconsciousness"]
walking independently; walking with auxiliary aids; walking with physical assistance; bedridden; unknown	['level of autonomy (mobility)']
y; n; unknown	['first episod of epilepsy', 'known history of epilepsy', 'history of allergy', 'history of recent trauma', 'pregnancy', 'history of drug abuse', 'history of alcohol abuse', 'anticoagulants or antiplatelet drug therapy', 'presence of prodromal symptoms', 'compliance with antiepileptic therapy', 'tloc during effort', 'tloc while supine', 'antiepileptic therapy already in place', 'drowsiness, confusion, disorientation as postcritical state', 'stiffness during the episode', 'drooling during the episode', 'tonic-clonic seizures', 'poly-pharmacological therapy', 'pale skin during the episode', 'eye deviation during the episode', 'diffuse vascular disease', 'neuropsychiatric disorders', 'presence of pacemaker', 'presence of defibrillator', 'cardio-pulmonary resuscitation', 'antihypertensive therapy', 'cardiovascular diseases', 'neurodegenerative diseases', 'peripheral neuropathy', 'immunosuppression', 'palliative care', 'situation description, like coughing, prolonged periods of straining, sudden abdominal pain, phlebotomy', 'problematic family context', 'need but absence of a caregiver', 'homelessness', 'living alone', 'chest pain', 'head or other districts trauma', 'tongue bite', 'agitation', 'foreign body in the airways', 'improvement of dyspnea', 'presence of dyspnea', 'dementia', 'general condition deterioration', 'ab ingestis pneumonia', 'further seizures in the ed', 'improvement of patient's conditions', 'neurologist consultation', 'ecg, any abnormality', 'ecg monitoring, any abnormality', 'eeg, any abnormality', 'thoracic ultrasound, any abnormalities', 'chest rx, any abnormalities', 'gastroscopy , any abnormalities', 'brain ct scan, any abnormality', 'brain mri, any abnormality', 'cardiac ultrasound, any abnormality', 'chest ct scan, any abnormality', 'pulmonary scintigraphy, any abnormality', 'abdomen ct scan, any abnormality', 'compression ultrasound (cus), any abnormality', 'performance of thoracentesis', 'administration of diuretics', 'administration of steroids', 'administration of bronchodilators', 'administration of oxygen/ventilation', 'blood transfusions', 'administration of fluids', 'heart failure', 'pneumonia', 'copd exacerbation', 'acute pulmonary edema', 'asthma exacerbation', 'respiratory failure', 'intoxication', 'covid 19', 'influenza and various infections', 'pneumothorax', 'situational syncope', 'epilepsy / epileptic seizure', 'pulmonary embolism', 'arrhythmia', 'cardiac tamponade', 'aortic dissection', 'acute coronary syndrome', 'hemorrhage', 'severe anemia', 'concussive head trauma']

Table 5: **List of the** 136 **Case Report Form tasks and their respective labels' space (*options*).** Several tasks share the same labels. The elements in *items* are used to define the classification tasks, and as targets s_T to quantify the relevance during the DECSELFMASK relevance attribution step.

In DEA per disp nea da 4 - 5 giorni . ĊĊ AP R :Ċ - IP B Ċ -port atore di 3 BP AC (199 5) con ev ol uzione ip oc inet ica
Ċ -re se zione gastr ica per ul c era Ċ -dis lip id emia Ċ -ex f um atore Ċ - FA perman ente Ċ -port atore di PM Ċ -st en osi
val v ola a ort ica sever a Ċ - Nel 202 0 F NA TC guid ata di les ione pol mon are lob are infer iore sn : AD K (in consider
azione delle com orb ilit Ãł cardi olog iche non indic azione chir urg ica n Ã¨ radi oter ap ica).Ċ A lug lio 202 1 in izio Chem
iot er apia , success ivamente s ott op osto a immun oter apia ; TC total body 5 / 202 2 : regress ione del vers amento ple ur
ico e degli add ens amenti pol mon ari , non progress ione di mal att ia , indic azione a solo follow up Ċ - 1 / 202 2 ric over o
per disp nea in vers amento ple ur ico bil ater ale , > a dx , e s osp etta pol mon ite da farm aci ĊĊ ter apia in cor so :Ċ p
ant orc , ind ac ater olo , bis op rol olo 1 . 25 , ap ix aban 2 , 5 mg 1 co sera (ult ima ass un zione i eri sera), a ml od
ip ina 1 / 2 , tor v ast , las ix 1 c x 2 ĊĊ non allerg ie ĊĊ Il paz iente vive da solo , seg uito da assist ente sociale N OME
_PERSON A NUM _TE LE F ONO Ċ V acc in azione anti COVID 19 III dose

Figure 3: **Impact of the DecSelfMask training on the relevance attributed by AttnLRP to the input for the task *"Presence of pacemaker Pacemaker"*.** Red represents higher relevance being assign to a specific token by the base model when compared to its DecSelfMask version, while green the opposite. While the base model primarily focuses on terms related to *dyspnea*, the DecSelfMask model shifts its attention toward mentions of "*PM*" (pacemaker), which is substantially more relevant to the target condition

APP : In PS cond otta da 118 . N OME _PERSON A alle 09 . 30 circa cad uta a terra a din am ica non nota av ven uta al
domic ilio , il P AREN TE non ha assist ito alla cad uta ma Ã¨ acc or so dopo aver sent ito il rum ore . La paz iente si era giÃł
al z ata dal letto . Rit rov ata vig ile dal P AREN TE . Tra uma del vol to contro lo sp ig olo di un mobile . Ċ All ' arr ivo del
118 paz iente a terra , in dec ub ito later ale sin istro . Paz iente ob n ub il ata , res pi ro spont ane o presente , G CS 13 .
Non apparent i deficit di l ato , pup ille apparent emente is oc or iche , non seg ni di m ors us nÃ© r il asc io sf inter ico . E
vid enza di s angu in amento a car ico della pir amide nas ale e F LC a liv ello nas ale e per ior bit aria dx . Som min istr ato
ac ido tran exam ico 1 g . All ' EC G rit mo da PM . Pos iz ion ato coll are cerv ic ale rig ido ed eff ett u ata immobil
izzazione su spin ale con rag no .ĊĊ AP R : Ċ - Mor bo di Alzheimer con grave deterior amento cogn it ivo , epis odi di ir requ
iet ez za e wandering , con aff acc end amento durante la not te quasi quot idian i ; de amb ul azione inst abile (vis ita di
control lo CDC D mol in ette 6 / 4 / 202 2)Ċ - 200 7 re se zione di ne op las ia della gu ancia de stra ; 201 8 inter vent
o di s br igli amento Ċ - P reg ress o im pi anto PM bic amer ale End urity 217 2 (norm of un zion ante a control lo del 14
/ 05)Ċ - Cr isi epile tt ica recent e (14 / 5 / 22) per cui av vi ata ter apia ant ie pile tt ica con fen obar b ital (r ifer ita
dal P AREN TE ade gu ata compliance). TC en cef alo neg ativa per les ioni a f ocol a io . EEG (2 / 8 / 22): alter azione
dell 'e lett rogen esi cort ic ale con as pet ti dis fun z ional i b item por ali e r ari gra fo element i par oss ist ici diff usi , inf
ici ato da arte f atti . Ċ - I pot i roid ismo ĊĊ Non altri P AREN TE oltre al P AREN TE . NON ha tut ore leg ale ne ' am min
istr atore di sost eg no . ĊĊ Non allerg ie note .ĊĊ V acc in ata COVID 19 4 dos i .Ċ Non apparent emente vacc in ata con anti
-t et an ica .ĊĊ Cont atti di r ifer imento (P AREN TE): NUM _TE LE F ONO

Figure 4: **Impact of the DecSelfMask training on the relevance attributed by AttnLRP to the input when generating the text *"Antiepileptic therapy already in place"*.** Red represents higher relevance to be assign to a specific token by the base model when compared to its DecSelfMask version, while green the opposite. Our model assigns much more importance to "fenobarbital", whichis an antiepileptic drug, and reduces it to Alzheimer (irrelevant here). Interestingly, relvenace is lowered for "Crisi epilettica" (i.e., epileptic crisis), suggesting that the model has learnt to move attention to plausible portions to the ones containing the actual information. Epileptic crisis is obviously linked to its therapy, but gives no factual knowledge about medications being in place already.

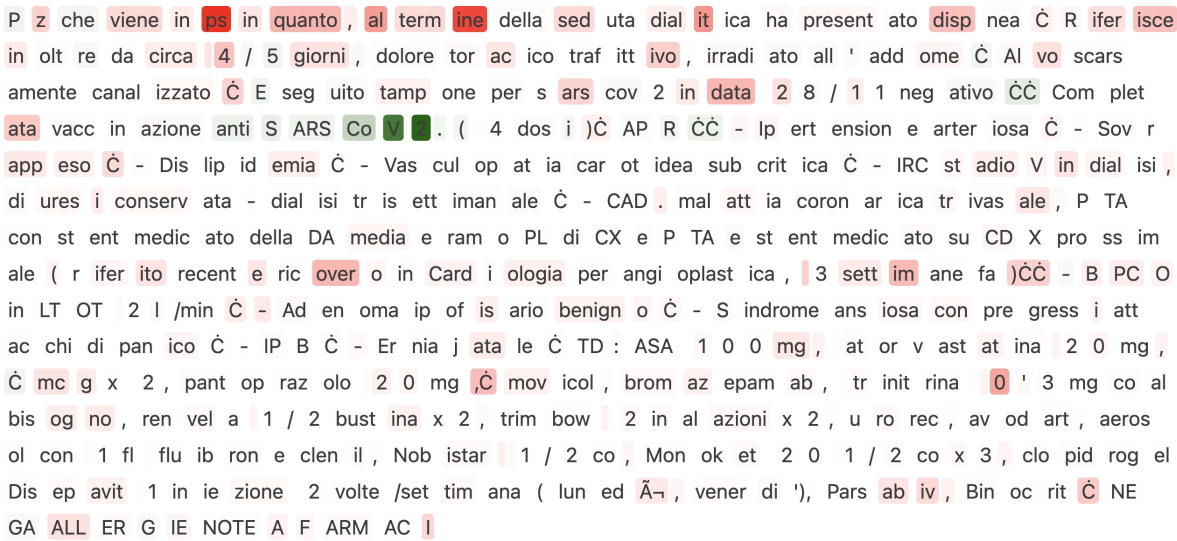

Figure 5: **Impact of the DecSelfMask training on the relevance attributed by AttnLRP to the input when generating the text *"SARS-CoV-2 swab test"*.** Red represents higher relevance to be assign to a specific token by the base model when compared to its DecSelfMask version, while green the opposite. Our model assigns much more importance to the token "Anti SARS CoV2", and reduces it to other tokens that are indeed not required by the target.

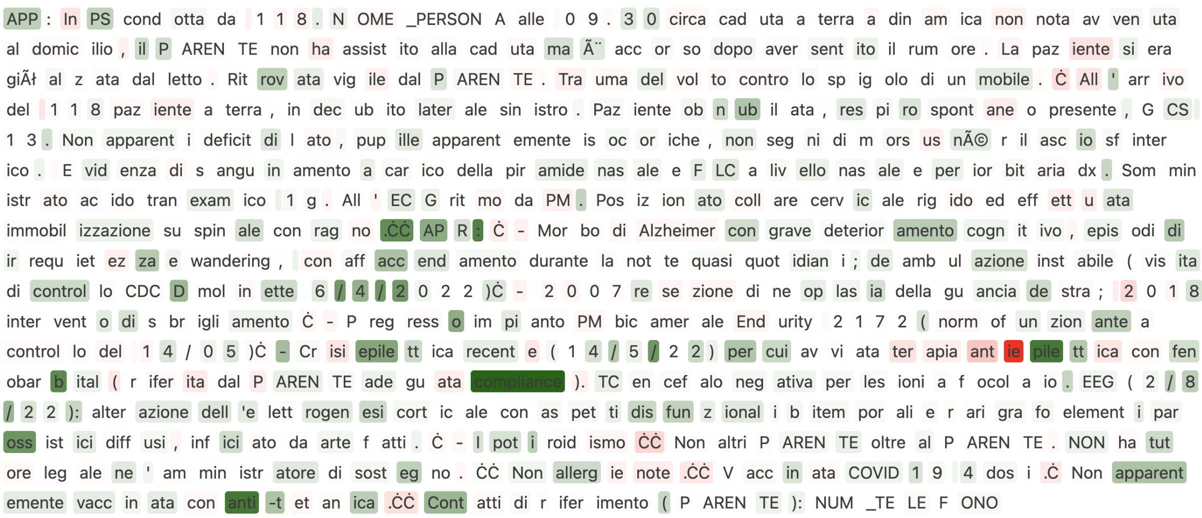

Figure 6: **Impact of the DecSelfMask training on the relevance attributed by AttnLRP to the input when generating the text *"Compliance with antiepileptic therapy"*.** Red represents higher relevance to be assign to a specific token by the base model when compared to its DecSelfMask version, while green the opposite. Our model assigns much more importance to the tokens "compliance", "fenobarbital", which are both relevant to the target.

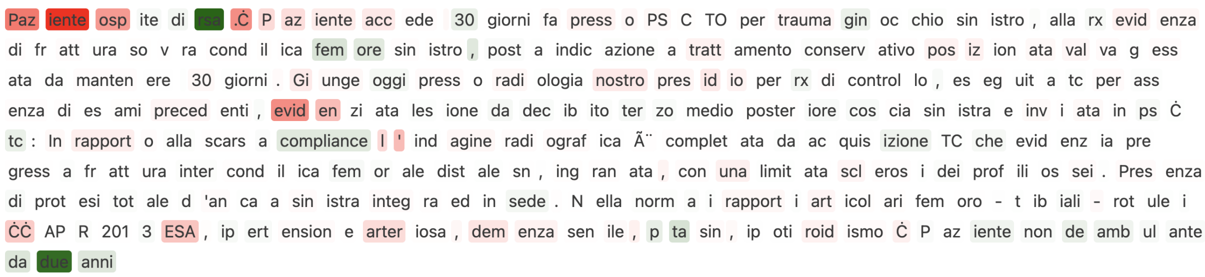

Figure 7: **Impact of the DecSelfMask training on the relevance attributed by AttnLRP to the input when generating the text *"level of autonomy"*.** Red represents higher relevance to be assign to a specific token by the base model when compared to its DecSelfMask version, while green the opposite. Our model assigns much more importance to the tokens "rsa" (i.e., care home) and "Paziente non deambula da due anni" (The patient has not been able to walk for two years).