# Modeling the Language of Holocaust Survivors' Testimony with Domain-Adapted Transformers

**Christopher Brückner**[1], **Jan Lehečka**[2], **Jan Švec**[2], **Pavel Pecina**[1]
[1]Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Prague, Czech Republic
[2]University of West Bohemia, Department of Cybernetics
Pilsen, Czech Republic
{bruckner,pecina}@ufal.mff.cuni.cz, {honzas,jlehecka}@kky.zcu.cz

**Abstract**

Documents related to the Holocaust increasingly move into the focus of Natural Language Processing research, including the digitization of written text, the automatic transcription of oral archives, and interpretive downstream tasks such as Named Entity Recognition. However, most modern language models are trained primarily on modern text, and thus struggle with historical language, historical entities, and domain-specific terminology. Furthermore, transcribed speech introduces challenges such as transcription errors, noise, filler words, and dialectal speech not often contained in textual datasets. We present XLM-RoBERTa-malach, a text encoder domain-adapted to oral testimonies of Holocaust survivors in seven languages. In addition to descriptions of the data acquisition via Automatic Speech Recognition, data augmentation via Machine Translation, and the continued pretraining of a state-of-the-art multilingual transformer, we evaluate the domain-adapted model on the Named Entity Recognition task. Experiments on this task show superior performance over the general-domain transformer in a multilingual domain-specific setting, including languages not seen during the domain adaptation.

**Keywords:** Language Modeling, Domain Adaptation, Holocaust Testimony, Speech Recognition

## 1. Introduction

With the advent of Transformers, Natural Language Processing (NLP) has made significant improvements in general-domain and domain-specific settings. However, most language models have been trained on modern text and do not generalize well to historical documents, which come with additional challenges, such as orthographic reforms, entity drift, noisy inputs, and a lack of resources (Ehrmann et al., 2023).

While language models domain-adapted to 18th to 20th century text have been shown to outperform general-domain models in the Named Entity Recognition (NER) downstream task in documents from the same time period (Schweter et al., 2022), and NER in Holocaust survivors' testimonies has become of interest (Dermentzi and Scheithauer, 2024), no language model adapted specifically to mid-20th century languages and Holocaust-related terminology does yet exist. With the increasing number of digitized documents, such a model becomes an interesting candidate to assist with the processing of large archives, at potentially better quality than the currently available domain-agnostic solutions.

An additional challenge introduced in this domain comes from the fact that many testimonies, especially in non-English languages, exist only in oral form, which is very different from written documentation and often requires Automatic Speech Recognition (ASR) technologies to make them accessible for further processing (Lehečka et al., 2023).

In this paper, we present XLM-RoBERTa-malach, a multilingual Transformer based on the XLM-RoBERTa architecture (Conneau et al., 2020), domain-adapted to Oral Holocaust Testimony. It is named after the Hebrew word for "angel", or Multilingual Access to Large Spoken ArCHives. The following sections describe the foundations of domain adaptation and Holocaust-related NLP, the acquisition and augmentation of training data via automatic speech recognition and machine translation, the training process, and finally, NER experiments in testimonies showing the outperformance of general-domain models.

While we are not able to publish the dataset itself due to the licensing of the source data, the domain-adapted model is available on Hugging Face[1] under the MIT license. The source code of the NER experiments is available on GitHub[2].

## 2. Related Work

### 2.1. Domain-Specific Language Models

Language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) are powerful

[1] https://huggingface.co/ufal/xlm-roberta-malach
[2] https://github.com/chbridges/malach-ner



*Proceedings of The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026*, pages 74–83
11 May 2026. 

and, even in the age of Large Language Models, resource-friendly text encoders popular for tasks that do not require language generation. Pretrained on large amounts of text, fine-tuning them on downstream tasks such as Named Entity Recognition (NER) allows them not only to solve the given tasks, but also to adapt their parameters to the language and domain present in the fine-tuning dataset. While the underlying language models have typically been pretrained on domain-agnostic text, it has been shown that models can achieve even better results if they have already been pretrained on text in the same domain as they will eventually be fine-tuned on.

For example, language models are typically pretrained primarily on modern text and do not generalize well to historical documents. Such texts are subject to language change in various dimensions, including changing spelling conventions, words losing or gaining additional meanings, and locations changing their names (Ehrmann et al., 2023). Within the HIPE-2022 shared task on NER in historical documents (Ehrmann et al., 2022), Historical Multilingual BERT (Schweter et al., 2022) has been pretrained from scratch on 19th- and 20th-century newspapers and outperformed other participating systems in multiple languages.

A different approach to domain adaptation is the continued pretraining, where an already pretrained model is further trained on the same training objective, but on new data. For example, Gururangan et al. (2020) continued the Masked Language Modeling pretraining of RoBERTa in different domains (biomedicine, computer science, news, and Amazon reviews) and achieved significantly improved results in domain-specific tasks such as relation and topic classification. XLM-RoBERTa, a highly multilingual transformer model still used in state-of-the-art NER architectures (Straková and Straka, 2025), has been additionally pretrained on parliamentary proceedings, outperforming the original general-domain model in sentiment analysis in the legal domain (Mochtak et al., 2024).

### 2.2. NLP for Testimonial Data

While Named Entity Recognition is a well-established task, its applications in speech are still limited, and language models not traditionally trained on spoken language struggle with this task (Caubrière et al., 2020; Yu et al., 2025). This poses a problem in the Holocaust domain, as large amounts of its documentation exist only in oral testimony. These testimonies are often not manually transcribed, which has led to the emergence and reliance on domain-specific Automatic Speech Recognition technologies (Lehečka et al., 2023).

First steps in NER in testimonial data have been taken by Anuradha Nanomi Arachchige et al. (2023), who labeled English testimonies from the United States Holocaust Memorial Museum[3], Fortunoff Video Archive[4], and the Wiener Holocaust Library[5] with a highly domain-specific and granular entity type ontology. In addition to the common entity types Person, Location, and Organization, it distinguishes between different spatial entities (Location, Geopolitical Entity, Ghetto, Camp, Street, River) and temporal entities (Time, Date, Event), as well as Military organizations, Warships, Spousal Relationships, and Languages. These distinctions can lead to ambiguities, as the types of toponyms can be context-dependent, e.g., "Czestochowa" can refer either to a city (Location) or a Camp. Baseline experiments show that general-domain language models outperform Historical Multilingual BERT (Schweter et al., 2022): While the Holocaust undoubtedly belongs into the historical domain, many testimonies have been recorded at the end of the 20th century, which most historical data predates.

The same testimonies have recently served as training data for the domain adaptation of an English language model, HoloBERT, which outperforms the general-domain BERT on some, but not all, entity types in this granular ontology (Anuradha et al., 2025).

A more recent, multilingual NER dataset in this domain is EHRI-NER (Dermentzi and Scheithauer, 2024), which is based on EHRI Online Editions[6] in 9 languages. EHRI-NER uses a smaller, but still domain-specific entity ontology, extending the standardized CoNLL format (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) with Dates, Camps, and Ghettos. However, since these Online Editions were not originally annotated for Named Entity Recognition, but for Entity Linking, EHRI-NER contains non-standard annotations. For instance, "father" is not a named entity, but can be tagged as Person, since the word can be linked to a specific entity within the context of the testimony it appears in. EHRI-NER has been published, split into languages[7] or into multilingual training/validation/test data[8].

## 3. Training Data

This section describes the acquisition and augmentation of a training corpus for the domain adaptation

[3] https://www.ushmm.org
[4] https://fortunoff.library.yale.edu
[5] https://www.testifyingtothetruth.co.uk
[6] https://www.ehri-project.eu/ehri-online-editions/
[7] https://github.com/EHRI/EHRI-NER
[8] https://huggingface.co/datasets/ehri-ner/ehri-ner-all

of a new multilingual language model.

### 3.1. Source Corpus

The Visual History Archive[9], maintained by the USC Shoah Foundation, is the largest existing archive of Holocaust testimonies. It comprises more than 55,000 video interviews with survivors and covers more than 30 languages. Founded in 1994, these interviews are rather recent and thus not subject to orthographic reforms or other significant evolutions of written language. However, beside terminology related to World War 2, Nazi Persecution, and Jewish Identity in these oral testimonies, survivors often refer to places by historical names that have fallen out of regular use.

USC has provided the video interviews in six languages, the distribution of which is shown in Table 1. In total, these 33,902 testimonies amount to more than 60,000 hours of MP4 files at a size of 27 TB.

| Language | Testimonies | Fraction |
|---|---|---|
| English | 28,457 | 83.94% |
| Polish | 1,521 | 4.49% |
| Hungarian | 1,369 | 4.04% |
| Dutch | 1,077 | 3.18% |
| German | 917 | 2.70% |
| Czech | 561 | 1.65% |
| Total | 33,902 | 100.00% |

Table 1: The distribution of languages in the available testimonies from the Visual History Archive.

### 3.2. Data Preparation

#### 3.2.1. Automatic Speech Recognition

The audio tracks from all 27 TB of MP4s have been extracted with FFmpeg to single-channel MP3s at a sampling rate of 16,000 Hz and a variable bitrate of 190-250 kbps. While this encoding is rather lossy, its quality is sufficient for automatic speech recognition (ASR), and it reduces the size of the data tremendously to only 1.55 TB, which helps significantly with data transfer and processing.

The ASR processing was performed using a self-hosted, containerized version of the UWebASR service (Švec et al., 2025)[10], deployed using Singularity containers in the MetaCentrum HPC infrastructure. The system supports the languages relevant to the Holocaust testimonies and utilizes two primary architectures: Wav2Vec 2.0 (Baevski et al., 2020) and the more recent Zipformer architecture (Yao et al., 2023). To efficiently manage long audio inputs typical of Holocaust testimonies, the engine employs a sliding window approach during transcription.

The training methodologies and decoding strategies differed between the two architectures. The Wav2Vec 2.0 models for Czech, Slovak, German, and English were first pre-trained on large-scale, unlabeled speech datasets (e.g., 80,000 hours for Czech and German, 20,000 hours for Slovak). These base models were then fine-tuned in a two-phase process: first on general-domain speech and subsequently on oral-history-style recordings (Lehečka et al., 2023). For Dutch, we utilized a publicly available Wav2Vec 2.0 model fine-tuned on the Corpus Gesproken Nederlands (CGN) dataset[11]. For decoding, the Wav2Vec 2.0 models employ Connectionist Temporal Classification (CTC) (Graves et al., 2006) over graphemes, integrated with an external language model to enhance linguistic context.

In contrast, the Zipformer models were trained using supervised learning directly on labeled datasets (see Table 2 for data sizes) and employ greedy CTC decoding over subword units (SentencePiece). The Zipformer architecture utilizes a modified Transformer encoder operating at multiple lower frame rates, enabling faster decoding and improved performance.

The architectures also cover different sets of languages. Wav2Vec 2.0 models were used for Czech, Slovak, German, English, and Dutch. The Zipformer architecture was applied to the same set (with the exception of Dutch) and further extended to include Polish and Hungarian. Table 2 shows the Word Error Rates (WER) for the languages matching our corpus across these architectures, compared against the Whisper-large-v3 baseline. The service supports flexible downstream processing by providing outputs in multiple formats, including plain text, WebVTT, JSON, and Transcriber XML.

Table 2 summarizes the labeled training data composition for each language, reporting both the total amount of supervised speech and the proportion originating from the oral-history domain. For nearly all languages, oral-history recordings constitute only a small fraction of the available labeled data, highlighting the severe scarcity of in-domain supervision and the resulting difficulty of the ASR task. Consequently, the evaluated models must rely heavily on cross-domain generalization rather than extensive domain-matched training. The only exception is German, for which substantially larger in-domain resources are available: manual annotations exist for approximately 900 German-language interviews, totaling nearly 2,000 hours, prepared by researchers from Freie Universität Berlin.[12]

[9] https://vha.usc.edu

[10] https://uwebasr.zcu.cz

[11] https://huggingface.co/GroNLP/wav2vec2-large-xlsr-53-ft-cgn

[12] Transcripts are publicly available at https://

| Language | Sup. Data [h] | Oral Hist. [h (%)] | Whisper v3 | Wav2Vec 2.0 | Zipformer |
|---|---|---|---|---|---|
| Czech | 6,000 | 106 (1.8%) | 19.1 | 8.5 | 7.1 |
| Slovak | 3,800 | 98 (2.6%) | 22.0 | 11.6 | 10.3 |
| German | 6,100 | 1,800 (30.0%) | 25.9 | 16.6 | 12.4 |
| English | 12,500 | 255 (2.0%) | 18.0 | 12.9 | 11.5 |
| Polish | 1,400 | 53 (3.9%) | 22.8 | – | 15.7 |
| Hungarian | 3,800 | 24 (0.6%) | 30.9 | – | 16.4 |

Table 2: Supervised training data size in hours [h] and Word Error Rates (WER) [%] of ASR architectures evaluated on oral history archives. Columns report the total amount of supervised data per language, the amount originating from the oral history domain (hours and proportion) and measured performance on the test split of the oral history dataset. A lower WER value indicates a better model. Whisper-large-v3 (Radford et al., 2022) serves as the general-domain baseline. We omit Dutch in this table as we had no labeled oral history data to fine-tune or evaluate the models for this language.

The ASR output is further processed through a post-processing pipeline for automatic punctuation and casing restoration. For English, German, Czech, and Slovak, we employed the approach described in Švec et al. (2021), using monolingual BERT-based predictors trained on CommonCrawl web text dumps to restore sentence boundaries, punctuation (full stop, comma, question mark), and proper casing. For the remaining languages (Polish, Hungarian, and Dutch), we utilized the `xlm-roberta_punctuation_fullstop_truecase` model[13] (Guhr et al., 2021). This step ensures that the resulting 3.1 GB of plain text is well-formatted for the subsequent domain adaptation of the corpus.

#### 3.2.2. Data Augmentation and Sampling

Due to the significantly skewed language distribution shown in Table 1, the produced text has been further machine-translated into all six present languages plus Danish to overcome data scarcity and create language-balanced training data. These translations have been created with MADLAD400-3B-MT[14], which has been shown to outperform comparable state-of-the-art models such as NLLB (NLLB Team et al., 2022) on mid- and high-resource languages at decreased inference time (Kudugunta et al., 2023; Lanz and Pecina, 2025). This includes the seven targeted languages.

The resulting balanced dataset has been tokenized with the XLM-RoBERTa tokenizer (Conneau et al., 2020), since this is the model architecture to be domain-adapted. Similar to the pretraining data of this architecture, the tokens have then been language-wise concatenated to single long strings and split into continuous, equally sized batches of 512 tokens. The final shorter batch in each language, which counts less than $10^{-7}$% of the total number of tokens, has been truncated.

Finally, 10% of batches per language are randomly sampled for a test set, and their tokens are statically masked with 15% probability. The remaining batches will be dynamically masked during the training, as in the Masked Language Model objective during the original XLM-RoBERTa pretraining.

Since 1/7 of the dataset is the output of different domain-specific and general-domain ASR models, and the remaining 6/7 are machine translations of the ASR output, this corpus can be considered 100% synthetic, although it is 100% a representation of real testimonies of Holocaust survivors. Due to possible errors and biases introduced by ASR artifacts and MT hallucinations, the corpus should not be used to train generative models, but only encoders for natural language understanding tasks such as NER.

### 3.3. Corpus Statistics

The created corpus has a total size of 4.9 billion tokens. The sizes per language and split are shown in Table 3. Although the same 33,902 testimonies are present in all seven languages, minus the truncated final batches, the numbers of tokens are not perfectly balanced across the languages. Instead, they represent how many tokens are required in each language to describe the same data.

| Language | Training | Test | Total |
|---|---|---|---|
| Czech | 637 | 71 | 708 |
| Danish | 612 | 68 | 680 |
| Dutch | 626 | 70 | 696 |
| English | 612 | 68 | 680 |
| German | 634 | 70 | 704 |
| Hungarian | 642 | 71 | 713 |
| Polish | 645 | 72 | 717 |
| Total | 4,407 | 490 | 4,897 |

Table 3: The VHA corpus size in **million [M]** tokens.

Since the test splits have been randomly sam-

`transcripts.vha.fu-berlin.de`.

[13] `https://huggingface.co/1-800-BAD-CODE/xlm-roberta_punctuation_fullstop_truecase`

[14] `https://huggingface.co/google/madlad400-3b-mt`

pled from each language individually, some data leakage has to be assumed: All test samples are likely seen during the training, albeit in different languages and with different positional embeddings. This can lead to an underestimated intrinsic perplexity of the language model. However, it does not affect extrinsic evaluation metrics on other datasets and downstream tasks.

In addition, we repurpose the full EHRI-NER dataset (Dermentzi and Scheithauer, 2024) to a second MLM test dataset by removing its annotations and applying the same tokenization and masking steps to it. We consider two variants of this test set: **EHRI-6** contains the six languages that overlap with our corpus (minus Danish), and **EHRI-9** additionally contains French, Slovak, and Yiddish, which our model does not see during the continued pre-training. The language distribution of this dataset is shown in Table 4. While significantly smaller in size (0.02%) and imbalanced, it is unbiased.

| | | Language | Tokens [k] |
|---|---|---|---|
| EHRI-9 (874) | EHRI-6 (713.5) | Czech | 195 |
| | | Dutch | 2.5 |
| | | English | 81 |
| | | German | 356 |
| | | Hungarian | 45 |
| | | Polish | 34 |
| | | French | 3.5 |
| | | Slovak | 6 |
| | | Yiddish | 151 |

Table 4: The size of the EHRI dataset in thousand [k] subword tokens. The total sizes of EHRI-6 and EHRI-9 are given in parentheses.

## 4. Domain Adaption

This section describes the domain adaption process and the internal evaluation of the resulting model on its pretraining objective.

### 4.1. Continued Pretraining Setup

We adapt the large-sized XLM-RoBERTa model[15] to the domain of Holocaust testimony by continuing its pretraining with the Masked Language Modeling objective on the produced VHA corpus. To do so, we replicate most of the hyperparameters reported Liu et al. (2019) for the original large-sized RoBERTa model[16], which uses Adam optimization (Kingma and Ba, 2017) with $\beta_1 = 0.9, \beta_2 = 0.98, \varepsilon =$ 1e-6, and 0.01 weight decay, a peak learning rate of $4e^{-4}$ that is warmed up for the first 6% of steps and then decays linearly, and an effective batch size of 8192 for 500k steps.

In contrast, we use the improved AdamW optimizer (Loshchilov and Hutter, 2019), decrease the peak learning rate to prevent overfitting, and train the model on 4 L40 GPUs with per-device batch size 8, using 64 gradient accumulation steps for an effective batch size of 2048 for 34k steps. These smaller parameters are in line with other domain-adapted RoBERTa and XLM-RoBERTa models trained on smaller corpora (Gururangan et al., 2020; Mochtak et al., 2024).

While the test set is already pre-masked, the training set is dynamically masked with the same probability of 15%. We experiment with the learning rates 2e-5 and 1e-5 and evaluate the model's cross-entropy after every 10% of total steps on the VHA test data and on the EHRI dataset. The training lasts approximately 25 hours for 1 pass through the whole dataset, after which the model is rolled back to the checkpoint with the smallest VHA cross-entropy.

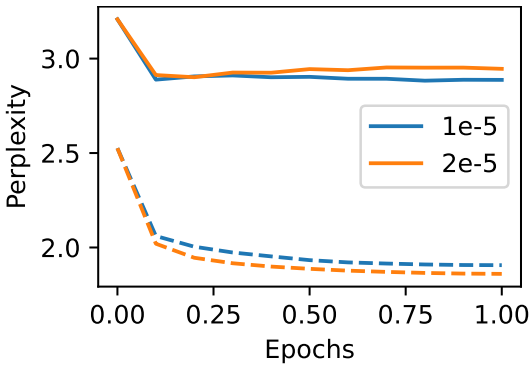


Figure 1: Perplexity of the model during 1 training epoch on the **VHA (dashed line)** and **EHRI-6 (solid line)** test datasets using peak learning rates of 1e-5 (blue) and 2e-5 (orange).

### 4.2. Resulting Model

Starting at 2.5257, the continued pretraining reduces the base-2 perplexity on the VHA data to 1.9064 (peak learning rate 1e-5) or 1.8603 (peak learning rate 2e-5). While Figure 1 indicates that neither model converges in 1 epoch, it also shows that the model with the greater learning rate starts overfitting to the automatically transcribed and translated VHA data 30% into the epoch, whereas the smaller learning rate keeps the perplexity on the EHRI-6 data stable at around 2.89. In both cases, the best checkpoint is the final checkpoint after 1 epoch.

Table 5 reports the perplexity of XLM-RoBERTa-large and both domain-adapted XLM-RoBERTa-

[15] https://huggingface.co/FacebookAI/xlm-roberta-large

[16] https://huggingface.co/FacebookAI/roberta-large

| Model | cs | de | en | fr* | hu | nl | pl | sk* | yi* |
|---|---|---|---|---|---|---|---|---|---|
| XLM-RoBERTa-large | 3.1553 | 3.4038 | 3.0588 | 2.0579 | 2.8928 | 2.9133 | 2.5284 | 2.6245 | **4.0217** |
| Malach 2e-5 | 2.8553 | 3.2277 | 2.9072 | *2.0966* | *2.9210* | *2.9404* | 2.4187 | *2.6592* | *5.3259* |
| Malach 1e-5 | **2.8023** | **3.1704** | **2.9022** | **2.0254** | **2.8285** | **2.8797** | **2.4003** | **2.5914** | *4.0910* |
| Support [k] | 195 | 356 | 81 | 3.5 | 45 | 2.5 | 34 | 6 | 151 |

Table 5: Base-2 perplexity of 3 models for 9 different languages in EHRI documents: The original XLM-RoBERTa-large checkpoint and ours, with two different peak learning rates. The last row shows the number of thousand [k] subwork tokens for each language. Languages marked with asterisks were not seen during the continued pretraining. Best scores are marked in **bold**, worse scores in *italics*.

malach variations in all 9 languages in the EHRI-9 corpus. Compared with the starting checkpoint, the peak learning rate of 2e-5 decreases the perplexity in only 4 of the seen languages, but increases it in Dutch, French, Hungarian, Slovak, and Yiddish. The increase from 4.0217 to 5.3259 in Yiddish, which is the only present language using a non-Latin alphabet, is particularly severe. In contrast, the peak learning rate of 1e-5 achieves the minimum perplexity in 8 languages, including 2 unseen ones. For Yiddish, the perplexity increases only to 4.0910. The greatest improvements can be observed in the Czech and German splits, which are also the greatest in size.

## 5. Named Entity Recognition

This section addresses the additional evaluation of the domain-adapted model on a domain-specific downstream task.

### 5.1. Experimental Setup

In addition to the internal evaluation of the domain-adapted models, we further evaluate them on the NER downstream task. The underlying dataset for this evaluation is the multilingual EHRI-NER (Dermentzi and Scheithauer, 2024), which has already served as an additional test dataset to measure the model perplexity on domain-specific data. EHRI-NER is annotated with the entity types Person, Organization, Location, Camp, Ghetto, and Date, in the languages Czech, Dutch, English, French, German, Hungarian, Polish, Slovak, and Yiddish.

`xlm-roberta-large-ehri-ner-all`[17], which has been fine-tuned on this dataset, serves as the baseline model. It is based on the same model as XLM-RoBERTa-malach, but without previous domain adaptation.

We replicate the training process of the baseline model on the published EHRI-NER training/validation/test splits: XLM-RoBERTa-malach is fine-tuned for 3 epochs using a learning rate of 3e-5, weight decay of 0.01, and a batch size of 16. The training is repeated 3 times using 3 different random seeds, so that not only the overall and tag-wise mean $F_1$ scores can be reported, but also their 95% confidence intervals.

[17] https://huggingface.co/ehri-ner/xlm-roberta-large-ehri-ner-all

### 5.2. Results

Mean $F_1$ scores and their 95% confidence intervals are shown in Table 6. Note that we were not able to exactly reproduce the results from Dermentzi and Scheithauer (2024) using their provided fine-tuned model and fixed test split, possibly due to differences in the processing and evaluation code.

The overall $F_1$ score does not increase significantly compared with the state of the art. Using a peak learning rate of 2e-5, it does not change at all; using a peak learning rate of 1e-5, it increases by 0.67% $F_1$ on average. More interesting are the differences per tag: While there is a slight decrease (less than 1%) for PER and LOC entities, the scores increase significantly (up to 5% on average) for the rarer, domain-specific CAMP and GHETTO entities.

While XLM-RoBERTa-malach pre-trained with a learning rate of 1e-5 tends to outperform the variant pre-trained with a learning rate of 2e-5 on the NER downstream task, its experimental results also come with increased variance, in particular with respect to ORG and GHETTO. Organizations appear to be generally difficult to predict in EHRI-NER, and ghetto examples are sparse. Furthermore, a typical error for all models is the confusion of camps, ghettos, and general locations.

In comparison, fine-tuning the English-centric domain-adapted HoloBERT (Anuradha et al., 2025) achieves an overall $F_1$ score of only 75%, with comparable scores only for the domain-specific entity types CAMP (72.00), GHETTO (82.00), and DATE (81.67). Its results are significantly worse for PER (77.00), ORG (56.67), and LOC (75.33).

## 6. Discussion

Overall, the model domain-adapted with a peak learning rate of 1e-5 processes domain-specific text better than the domain-agnostic model and the 2e-5 variant, in terms of internal metrics (Masked Language Modeling and perplexity) and external

| | EHRI | Malach 1e-5 | Malach 2e-5 |
|---|---|---|---|
| PER | **87.00** | 86.67 ± 1.43 | 85.67 ± 1.43 |
| ORG | 63.00 | 64.33 ± 6.25 | **64.33 ± 1.43** |
| LOC | **82.00** | 81.67 ± 1.43 | 81.67 ± 1.43 |
| CAMP | 70.00 | **75.00 ± 2.48** | 73.33 ± 2.87 |
| GHETTO | 80.00 | **85.00 ± 6.57** | 84.67 ± 1.43 |
| DATE | 84.00 | **85.00 ± 4.30** | 84.67 ± 3.79 |
| Overall | 81.00 | **81.67 ± 1.43** | 81.00 ± 0.00 |

Table 6: Mean micro $F_1$ scores (%) and their 95% confidence intervals on the EHRI-NER dataset. The compared models are `xlm-roberta-large-ehri-ner-all` (Dermentzi and Scheithauer, 2024) and our XLM-RoBERTa-malach, pre-trained with peak learning rates 1e-5 and 2e-5. Despite the variance of individual tags, the last model achieves the same overall score across 3 experiments.

metrics (Named Entity Recognition and $F_1$ scores). The perplexity decreases even for two languages unseen during the continued pretraining, which suggests that the model has learned genuine domain-specific representations, rather than simply memorising language patterns from the training data. However, the perplexity slightly increases for Yiddish, which is the only present language not using the Latin alphabet. Given the relevance of Yiddish in this domain, the increased perplexity is unfortunate, and additional data in Yiddish is required to tackle this issue. Such data can be generated via ASR (Marmor et al., 2025); however, machine translation from English to Yiddish is often of low quality (Kudugunta et al., 2023), and its suitability for data augmentation has to be further investigated.

In the NER task on multilingual written testimony, the performance noticeably increases on domain-specific entities, while the overall performance improves only slightly. This is because more general entities, such as people, occur much more frequently than camps and ghettos, which are of particular interest when extracting entities from testimonies. Slight degradations can be observed for PER and LOC entities. The latter one can be explained by the occasional ambiguity of LOC, CAMP, and GHETTO.

Although adapted exclusively to speech, produced by automatic speech recognition and machine translation, XLM-RoBERTA-malach is an interesting candidate for the processing of oral and written testimonies in multilingual settings. Its vastly improved performance in multilingual NER over HoloBERT (Anuradha et al., 2025) outlines the importance of multilingual pretraining in this domain. However, the model has not been evaluated on downstream tasks in speech data due to the limited availability of annotated corpora. In particular, the model has been trained on Danish, but no annotated domain-specific Danish text is yet available.

## 7. Conclusion and Future Work

We presented XLM-RoBERTa-malach, a variation of the large-sized multilingual XLM-RoBERTa model, domain-adapted to oral testimonies of Holocaust survivors. These testimonies have been produced via Automatic Speech Recognition of video testimonies in 6 languages from the Visual History Archive, the largest available archive of testimonies, and the resulting corpus has been further augmented via Machine Translation to add a seventh language, tackle data scarcity in six of the seven languages, and balance out the language proportions.

Although based on real testimonies, the resulting corpus can be considered synthetic, and both steps in the corpus creation can create errors and biases. Despite these issues, the continued pretraining on the Masked Language Modeling objective decreased the model perplexity not only on the speech-based training data, but also on written testimonies in 8 languages using the Latin alphabet: Czech, Dutch, English, French, German, Hungarian, Polish, and Slovak. Notably, French and Slovak were not seen during the domain adaptation, whereas the model has been additionally adapted to Danish. On the other hand, the model perplexity slightly increased on Yiddish, which is based on the Hebrew alphabet.

In the same languages, XLM-RoBERTa-malach outperforms its non-adapted variant XLM-RoBERTa-large on the Named Entity Recognition task in written testimonies. While it handles the frequent, general-domain entity type Person slightly worse, it exhibits significant improvements in the recognition of domain-specific entities, namely ghettos, camps, and numerical dates. Overall, the domain-adapted model appears to be an interesting baseline for NLP tasks in this domain. A very small learning rate, even smaller than the learning rate used during the NER fine-tuning (1e-5 < 3e-5), has proven to be beneficial in the domain adaptation of this model.

In the future, XLM-RoBERTa-malach should be evaluated in additional domain-specific downstream tasks, including NER in speech, such as the new MalachNER dataset (Brückner et al., 2026). The model itself can be improved in several ways:

- More language can be included in the domain adaptation corpus. For example, the used ASR system additionally supports Croatian and Serbian. Languages with non-Latin alphabets often spoken by survivors, e.g., Yiddish, Hebrew, Russian, and Ukrainian, can be added for better cross-lingual generalization. SOTA ASR models for Yiddish and Hebrew are available (Marmor et al., 2025), and further languages can be added via machine translation.

- In addition to speech transcripts, manual transcripts and written testimonies can be added to the training data to cover a larger variance of language and named entities.

- The possible data leakage in the model training, which affects the internal evaluation and convergence criteria, can be tackled by separating the testimonies in the training and test splits more clearly. I.e., no translations of training samples should appear in the test data, and vice versa.

Furthermore, the domain adaptation corpus can be used as a parallel corpus to train cross-lingual sentence embeddings (Reimers and Gurevych, 2019; Feng et al., 2022) for sentence similarity tasks, such as document retrieval and sequence classification.

## 8. Ethics Statement

Holocaust testimony is a sensitive domain and should always be handled with consideration. The automatic speech recognition and machine translation used to produce the training corpus may introduce errors that should not be present in published data related to the Holocaust. The domain-adapted model is an encoder-only model to be used for downstream tasks such as Named Entity Recognition, which mitigates the risk of reproducing biases often seen in causal language modeling, i.e., in autoregressive or diffusion models for text generation. However, this does not prevent the model entirely from misuse: We emphasize that results produced with XLM-RoBERTa-malach should still be carefully validated, e.g., before automatically processing archival material.

## 9. Limitations

The corpus used for the domain adaptation cannot be published, as the processed video data is licensed only for use within the project this research has been conducted in. The published domain-adapted model has only been evaluated on one downstream task, as, to our knowledge, no more annotated data in this domain is currently openly available.

## 10. Acknowledgements

This project is funded by the European Union's Horizon Europe research and innovation programme under grant agreement No. 101061016.

Disclaimer: Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic. The work described herein has also been using services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

This research was partially supported by Charles University, project GA UK No. 380126 and SVV project number 260 821.

## 11. Bibliographical References

Isuri Anuradha, Le An Ha, and Ruslan Mitkov. 2025. HoloBERT: Pre-trained transformer model for historical narratives. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 105–110, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Isuri Anuradha Nanomi Arachchige, Le Ha, Ruslan Mitkov, and Johannes-Dieter Steinert. 2023. Enhancing named entity recognition for holocaust testimonies through pseudo labelling and transformer-based models. In *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*, HIP '23, page 85–90, New York, NY, USA. Association for Computing Machinery.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Christopher Brückner, Karin Roginer Hofmeister, Jiří Kocián, and Pavel Pecina. 2026. From oral history to structured data: The MalachNER dataset. In *Proceedings of the Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026*, Palma de Mallorca, Spain. ELRA.

Antoine Caubrière, Sophie Rosset, Yannick Estève, Antoine Laurent, and Emmanuel Morin. 2020. Where are we in named entity recognition from speech? In *Proceedings of the Twelfth Language Resources and Evaluation Conference*,

pages 4514–4520, Marseille, France. European Language Resources Association.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Maria Dermentzi and Hugo Scheithauer. 2024. Repurposing holocaust-related digital scholarly editions to develop multilingual domain-specific named entity recognition tools. In *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*, pages 18–28, Torino, Italia. ELRA and ICCL.

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.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.*, 56(2).

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide. 2022. Extended overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, volume 3180. CEUR-WS.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. 2021. Fullstop: Multilingual deep models for punctuation prediction.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36:67284–67296.

Vojtech Lanz and Pavel Pecina. 2025. When multilingual models compete with monolingual domain-specific models in clinical question answering. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 69–82, Albuquerque, New Mexico. Association for Computational Linguistics.

Jan Lehečka, Jan Švec, Josef V. Psutka, and Pavel Ircing. 2023. Transformer-based Speech Recognition Models for Oral History Archives in English, German, and Czech. In *Interspeech 2023*, pages 201–205.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Yanir Marmor, Yair Lifshitz, Yoad Snapir, and Kinneret Misgav. 2025. Building an accurate open-source hebrew asr system through crowdsourcing. In *Proc. Interspeech 2025*, pages 723–727.

Michal Mochtak, Peter Rupnik, and Nikola Ljubešić. 2024. The ParlaSent multilingual training dataset for sentiment identification in parliamentary proceedings. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*

*(LREC-COLING 2024)*, pages 16024–16036, Torino, Italia. ELRA and ICCL.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stefan Schweter, Luisa März, Katharina Schmid, and Erion Çano. 2022. hmbert: Historical multilingual language models for named entity recognition. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 1109–1129. CEUR-WS.org.

Jana Straková and Milan Straka. 2025. NameTag 3: A tool and a service for multilingual/multitagset NER. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–39, Vienna, Austria. Association for Computational Linguistics.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, Fangjun Kuang, Yifan Yang, Zengrui Jin, Long Lin, and Daniel Povey. 2023. Zipformer: A faster and better encoder for automatic speech recognition. *arXiv preprint arXiv:2310.11230*.

Jiawei Yu, Xiang Geng, Yuang Li, Mengxin Ren, Wei Tang, Jiahuan Li, Zhibin Lan, Min Zhang, Hao Yang, Shujian Huang, and Jinsong Su. 2025. "i've heard of you!": Generate spoken named entity recognition data for unseen entities. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Jan Švec, Jan Lehečka, and Pavel Ircing. 2025. Current State of the UWebASR - Web-Based ASR Service for Czech, Slovak, German, and English. In *CLARIN Annual Conference Proceedings 2025*, page 95.

Jan Švec, Jan Lehečka, Luboš Šmídl, and Pavel Ircing. 2021. Transformer-Based Automatic Punctuation Prediction and Word Casing Reconstruction of the ASR Output. In *Text, Speech, and Dialogue. TSD 2021*, pages 86–94.