# Fine-tuning BERT Models on Demand for Information Systems Explained Using Training Data from Pre-modern Arabic

Thomas Asselborn[1], Sylvia Melzer[1,2], Said Aljoumani[2], Magnus Bender[1], Florian Andreas Marwitz[1], Konrad Hirschler[2] and Ralf Möller[1]

[1] *University of Lübeck, Institute of Information Systems, Ratzeburger Allee 160, 23562 Lübeck, Germany*
[2] *Universität Hamburg, Centre for the Study of Manuscript Cultures, Warburgstraße 26, 20354 Hamburg, Germany*

Abstract

Humanities scholars can use Large Language Models (LLMs) to simplify their work to analyse texts or recognise patterns in data. However, if a domain-specific problem needs to be solved, the existing models need to be fine-tuned. In the humanities, there is less training data available for fine-tuning, but there are increasing information systems with research data in that can be used for this purpose. To use data from information systems, humanities researchers need to be able to transform research data from information systems to training data so that it is suitable for training LLMs. However, assigning the correct labels (e.g., person, location, date) to the data can be challenging. This article describes how to fine-tune BERT models on demand for information systems explained using training data from pre-modern Arabic. The result we have achieved is that all archived research data can be used in a research data repository for fine-tuning models in short time and in a simplified way, i.e., without being an IT expert. Since the data was available in canonical form, it was possible to specify which fields could be assigned to which label by means of a manifest file. The results we obtained show that the fine-tuning process can be done in just a few minutes using a sample dataset and BERT. The Fine-tuning on Demand (FToD) process identified names of people, places, or dates that could not be recognised by the pre-trained model.

**Keywords**

Fine-tuning on demand, BERT, pre-modern Arabic, manifest file

*Humanities-Centred AI (CHAI), 3rd Workshop at the 46th German Conference on Artificial Intelligence, September 26, 2023, Berlin, Germany*

thomas.asselborn@uni-luebeck.de (T. Asselborn); sylvia.melzer@uni-hamburg.de (S. Melzer); saidaljomani@gmail.com (S. Aljoumani); bender@ifis.uni-luebeck.de (M. Bender); f.marwitz@uni-luebeck.de (F. A. Marwitz); konrad.hirschler@uni-hamburg.de (K. Hirschler); moeller@uni-luebeck.de (R. Möller)
https://www.ifis.uni-luebeck.de/index.php?id=asselborn (T. Asselborn); https://www.csmc.uni-hamburg.de/about/people/melzer.html (S. Melzer); https://www.csmc.uni-hamburg.de/about/people/aljoumani.html (S. Aljoumani); https://www.ifis.uni-luebeck.de/~bender/ (M. Bender); https://www.ifis.uni-luebeck.de/index.php?id=marwitz (F. A. Marwitz); https://www.aai.uni-hamburg.de/voror/personen/hirschler.html (K. Hirschler); https://www.ifis.uni-luebeck.de/~moeller/ (R. Möller)
0009-0005-3011-7626 (T. Asselborn); 0000-0002-0144-5429 (S. Melzer); 0009-0004-8306-8621 (S. Aljoumani); 0000-0002-1854-225X (M. Bender); 0000-0002-9683-5250 (F. A. Marwitz); 0000-0002-6012-7711 (K. Hirschler); 0000-0002-1174-3323 (R. Möller)

# 1. Introduction

Large Language Models (LLMs) are a subfield of Natural Language Processing (NLP) and are based on transformer architecture neural networks that use deep learning algorithms [1, 2, 3]. They are pre-trained on huge amounts of text and fine-tuned for specific tasks. Some of the most popular LLMs are BERT (Bidirectional Encoder Representations from Transformers) [4], GPT-3 (Generative Pre-trained Transformer 3) [5], and T5 (Text-to-Text Transfer Transformer) [6]. BERT is an LLM specifically designed for text processing and is capable of modelling semantic relationships between words and sentences. It is used for various applications, such as automatic text summarisation, question-answering generation, and sentiment analysis. Although BERT has been trained with 3300M data from BooksCorpus and English Wikipedia [4], there are limitations to the application in specific domains. The limitations arise from the fact that the validity of a design is not equally applicable to everything. In Arab countries, for example, the names with first name and surname are adapted to the Western naming pattern (e.g. first name: *Hamza*, surname: *ibn Omar ibn Mustafa*). In the early Arabic texts as much as in texts from other world regions, however, there are different naming pattern. The name also included, for example, membership of a tribe, the origin of a place, or even the affiliation to a legal school or occupational title. A place name can thus also be part of a person's name and should therefore be labelled as such. It also happens that places are paraphrased, such as *the Golden Mosque*, *the Glittering School* or *the Big Tower*. The indication of a date is also different from the Gregorian calendar, as can be seen from the following sentence: *"Hamza ibn Omar ibn Mustafa went to Marrakesh on the ninth day of the month of Shawwāl."*.

The different date representations can be found in the collection of so-called audition certificates written in pre-modern Arabic. Audition certificates are notes on a written artefact (such as a codex, a scroll, or a sheet of paper) to document the authorised transfer of a text from the book from teacher(s) to student(s). At the Centre of Manuscript Cultures (CSMC) at the Universität Hamburg, research data was collected from over 3,000 audition certificates. A total of over 50,000 annotations were made in the texts of the audition certificates. These annotations are persons, places, and dates as well as their role in the reading session. Annotations and other research data were stored in an information system called Audition Certificate Platform (ACP)[1].

At the CSMC are about 13 other information systems which represent research data in different, project-specific data models. Each project has therefore collected and stored project-specific data. It is exactly this data that can now be used as fine-tuning to train the LLMs to the needs of the projects.

Considering that humanities scholars involved in a new project aim to assess textual artefacts in terms of person names, locations, dates, or other types of entities, and lack their own specific fine-tuning data, they have the option to utilize fine-tuning data obtained from other projects. Looking outside the CSMC, there are research data repositories like the Research Data Repository (RDR) or Zenodo that also have lots of archived data that could be used for fine-tuning.

Humanities scholars, as well as other researchers without specific IT knowledge, are already producing labeled data, i.e., texts with tags or categories assigned to specific words or sentences indicating meaning or relevance. These data might however be in a format that can not be

[1]https://heurist.fdm.uni-hamburg.de/html/heurist/?db=CSMC_UWA_BibliothecaArabica&website&id=408721

used directly to fine-tune a LLM, e.g., in Microsoft Word docx. In order to properly use the data, certain pre-processing steps might be needed which can be time consuming and hard to understand, requiring an IT expert to do the task.

Various tools and libraries are available for the process of building a fine-tuned model. However, the biggest challenge is to empower scholars to use them without requiring the assistance of an IT expert. This would enable scholars to build their own fine-tuned models using project-specific data archived in research data repositories. Therefore, an approach is needed to allow scholars to build fine-tuned models intuitively and independently. To address this need, we have developed the Fine-Tuning on Demand (FToD) approach, which enables scholars to build project-specific models on demand in just a few seconds and with minimal resources. In this article, we explain the FToD approach using training data from pre-modern Arabic. The FToD approach can also be applied to other data from different domains, making it universally applicable in the humanities as well as in other fields.

In many cases, using custom models for fine-tuning can improve results, but this is not always the case [7]. After we labelled the places described, such as *the Golden Mosque* as locations, the results became worse as places like *Baghdad* were no longer identified as locations. At this point, experts would have to evaluate the result and repeat the fine-tuning process.

The main contribution of this paper is to define the FToD process to reduce the time needed to pre-process a dataset to fine-tune a BERT model to a specific task. The users of the system do not need to know the specifics of the BERT models and the libraries used.

The results we obtained show that the fine-tuning process can be done in just a few minutes using a sample dataset and BERT. The FToD process identified names of people, places or dates that could not be recognised by the pre-trained model. In the end, we provide an extended outlook to ChatHA, a Humanities Aligned Chatbot trained on the same dataset as BERT.

## 1.1. Related Work

Various Data Management Plans (DMPs) have the focus on archiving research data for a long time and making it accessible to other researchers. To fulfill the requirements of funding programs or regulations, tools and research data repositories (RDRs) have been developed. These tools provide a structured description of the data, eliminating the need for individual researchers to handle it themselves. As a result, RDRs are well-suited for data archiving. Users can reuse this archived data to train LLMs, for example.

Scholars can use LLMs to simplify their work to analyze texts or recognize patterns in data. However, if a domain-specific problem needs to be solved, the existing models need to be fine-tuned. There is limited information on the application of fine-tuning large language models in the humanities. However, a multilingual transformer model called CAMeLBERT [8] for segmenting Arabic text without punctuation is employed. The results showed that the proposed model outperformed other models in terms of accuracy and demonstrate the potential of fine-tuning large language models in the humanities for various tasks. However, further research is needed to explore the full potential of these models in the humanities.

In [9], the survey of natural language processing (NLP) approaches presents recent work that uses large language models to solve NLP tasks via pre-training and fine-tuning, prompting and other techniques.

## 2. Audition Certificates

Audition certificates are a salient feature of Arabic manuscript cultures. They are notes written on an artefact to document the authorised transmission of the text from teacher(s) to student(s). In concrete terms, the text was read out aloud (by the teacher or one of the students) and at the end of the reading session one of the members of this reading group added the audition certificate to the book. By virtue of their participation all students now had the right to act as teacher in future reading sessions.

Audition certificates can include: the name of the teacher(s), the name of the student(s), the name of the reader, the name of the writer of the certificate, the name of the book's owner, the date of the reading, the place of the reading, and many other data.

One example of an audition certificate, both in original Arabic script and translation, is presented in the following. The example was taken from the manuscript Bibliothèque nationale de France, arabe 708, fol. 38v and can be found as a digital copy online[2] as well as in a Heurist database of the Universität Hamburg[3]. Persons are written in green while dates are pink and locations blue.

Original Arabic text:

قرأت جميع هذا الجزؤ وهو الثاني من كتاب السنن لأبي داود وجميع الجزء الثالث بعده على الشيخين الأجلين الإمام العالم الحافظ الفاضل قطب الدين أبي بكر محمد بن الشيخ القدوة أبي العباس أحمد بن علي القسطلاني بسنده المذكور في طبقة السماع في الجزء الأول والفاضل المسند شهاب الدين أبي الفضل عبد الرحيم بن يوسف ابن يحيى الدمشقي الشافعي عرف بابن خطيب المزة بسماعه من طبرزد فسمعهما السادة الأجلاء زين الدين أبو العباس أحمد وصدر الدين أبو الخير عبد البر ولدا سيدنا قاضي القضاة تقي الدين مفتي المسلمين أبي عبد الله محمد بن الحسين بن رزين الشافعي وولد أختهما فخر الدين أبو عمرو عثمان بن شمس الدين محمد وشرف الدين أبو العباس أحمد وبهاء الدين أبو البركات عبد الحق ولدا الشيخ قطب الدين ابن القسطلاني وابن أخيهما نور الدين علي بن أمين الدين أبي المعالي محمد ومحمد بن أحمد بن علي والفقيه الفاضل شمس الدين محمد بن أبي القاسم بن عبد [السلام] ابن جميل الربعي وشمس الدين خليل بن بدران بن خليل الحلبي الصوفي وهو مثبت أسماء الجماعة في الأصل وصح ذلك وثبت في يوم الأربعاء تاسع جمادى الأولى من سنة سبع وسبعين وستمائة وأجاز المسمعان لي وللمذكورين جميع ما يجوز له [كذا] روايته بدار الحديث الكاملية بالقاهرة كتبه الحسن بن علي بن عيسى بن الحسن بن علي اللخمي والحمد لله وحده وصلواته على سيدنا محمد وآله وصحبه وسلامه.

**Translation** "I have read all of this part, and it is the second from the book Al-Sunan by Abu Dawud, and all of the subsequent third part under the authority of the two eminent sheikhs, the scholar, the hafiz, the virtuous Qutb al-Din Abi Bakr Muhammad ibn al-Sheikh al-Qudwa Abi al-Abbas Ahmad ibn Ali al-Qastalani according his chain of transmission mentioned in the audition certificate that is in the first part, and the virtuous Musnad Shihab al-Din Abi al-Fadl Abd al-Rahim ibn Yusuf ibn Yahya al-Dimashqi al-Shafi'i known as Ibn al-Khatib al-Mizza by virtue of the authorised transmission from Tabarzad, distinguished masters Zain al-Din Abu al-Abbas Ahmad and Sadr al-Din Abu al-Khair Abd al-Barr the sons of our Master the Chief

[2] https://gallica.bnf.fr/ark:/12148/btv1b11000356s/f19.item.r=%22arabe%20708%22.zoom#

[3] https://heurist.fdm.uni-hamburg.de/html/heurist/?db=CSMC_UWA_BibliothecaArabica&w=a&q=t%3A101%20708%2016

Judge Taqi al-Din the mufti of the Muslims Abi Abd Allah Muhammad ibn al-Husain ibn Ruzayq al-Shafi'i and their nephew Fakhr al-Din. Abu Amr Othman ibn Shams Al-Din Muhammad and Sharaf Al-Din Abu Al-Abbas Ahmed and Baha' Al-Din Abu Al-Barakat Abdul-Haq, the sons of Sheikh Qutbuddin ibn Al-Qastalani and their nephew Nur Al-Din Ali ibn Amin Al-Din Abi Al-Ma'ali Muhammad and Muhammad ibn Ahmad ibn Ali and the virtuous jurist Shams Al-Din Muhammad ibn Abi Al-Qasim ibn Abd [al-Salam] ibn Jamil al-Raba'i and Shams al-Din Khalil ibn Badran ibn Khalil al-Halabi al-Sufi, and he established the names of the group in the original, and that was confirmed on Wednesday, the ninth of Jumada al-Awwal of the year seventy-seven and six hundred. The two authorised teachers gave me and those mentioned the authority to transmit all that had been authorised to them. [This reading took place] in the Dar al-Hadith al-Kamiliyah in Cairo, written by Al-Hassan ibn Ali ibn Isa ibn Al-Hassan ibn Ali Al-Lakhmi."

### 2.1. Audition Certificate Platform

Audition certificates are the result of complex documentary practices and they were written by highly specialised communities. We have collected all the research data about the audition certificated into a database. On top of the database we built an information system which is called Audition Certificate Platform.

ACP was implemented as an information system using the database management tool Heurist[4]. The objective is to analyze the collected research data using Heurist and make it accessible to users through a website. Once the project is completed, the information system can be exported in formats such as CSV, XML, or JSON and archived in a RDR. The archived data is now available as training data.

## 3. Fine-Tuning on Demand

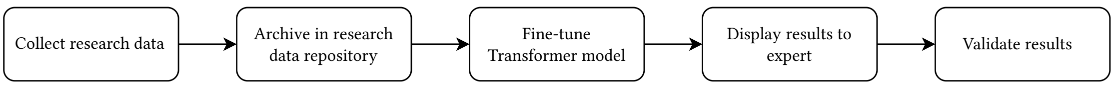


**Figure 1:** The process Fine-Tuning on Demand (FToD) has five steps: 1. collect research data, 2. archive in research data repository, 3. fine-tune Transformer model, 4. display results to expert, and 5. validate results.

The FToD process is presented in Figure 1. FToD consists of five steps. As a first step, researchers collect research data which can be stored in formats such as JSON, XML, CSV, or in a database. Once the researcher has collected all of their data, they will be archived in a research data repository together with a manifest file. We use the manifest file which is called Metadata Encoding & Transmission Standard (METS) file.

METS is a standard that enables the exchange of digitized documents between cultural heritage institutions and is an XML schema for the creation of digital objects. A digital object can consist of one or more digital files, which can be in different formats and describe a detailed

[4]https://heurist.fdm.uni-hamburg.de/index_en.html

internal structure. A METS-file consists of seven major sections. Here, an excerpt is presented of the fields important for FToD. Other fields are described in the standard [5].

- METS Header: The METS Header is a section within the METS-file that contains metadata about the METS-file itself. It includes information such as the creator and editor of the file.
- File Section: The file section in METS lists all the files that make up the digital object. It includes <file> elements that can be grouped within <fileGrp> elements to organize the files based on different versions or categories of the object. Additionally, it includes functions, for example, what data is to be harvested (e.g. JSON file) or which process is to be executed; in this case FToD. The function is represented with the attribute "USE", with the assignment USE="json" or USE="FToD."
- Structural Map: The structural map is a crucial component of a METS-file. It outlines the hierarchical structure of the digital library object, defining the relationships between different elements. It also links the elements to the corresponding content files and metadata. The Structural Map is built in a tree-like structure using multiple nested DIV elements, with the root DIV element containing all other DIV elements. The DIV elements directly under the root DIV element represent the possible models for training (e.g. BERT), and within the DIV elements of the models. The DIV element contains a Filepointer element that points to the File element of the research data file (e.g. JSON).
- As an extension of Tilp's METS Generator (see [10]), we also use the Structmap element to do the label assignment for the FToD process. Each DIV is assigned all the labels that are entered as input in BERT. The Filepointers would then represent the field names. Several Filepointers can be specified. Thus, the fields "Region" and "Territory" can be labelled with "Place."

METS-files can also be used for other processes such as the Databasing on Demand (DBoD) processes [11]. With the help of a METS-file, data stored in, e.g., a CSV-file and archived in a research data repository (RDR) are transferred to a database at the press of a button. (The button was added to the RDR in a prototype implementation.) More information on the practical application of the METS file and the DBoD process is described in Stahl's bachelor thesis [12].

Once everything is archived in the repository, a registered and authorised user will see a button saying "Train my model" (see Figure 2). Train my model was chosen in favour over fine-tune my model because it may be easier to understand for non-IT experts. Once the process has being started, a screen is shown to the user asking for patience (see Figure 3). Since the fine-tuning process might take a long time, depending on the dataset, hardware configuration, specific task, etc., the user will additionally receive an e-mail once the process has finished.

After fine-tuning is completed, the user is being redirected to a screen where they are able to verify the results (see Figure 4). This screen is shown when the BERT named entity recognition task has been chosen. Verification screens for other tasks will be implemented in the future as necessary. Additionally, the user also receives an email with a link to this screen. In some circumstances, the process can take a little longer, so we have included a function in a prototype implementation that informs the user by email when the process has been completed successfully. More details on the implementation of a notification function are described in [12].

[5]https://www.loc.gov/standards/mets/METSOverview.v2.html

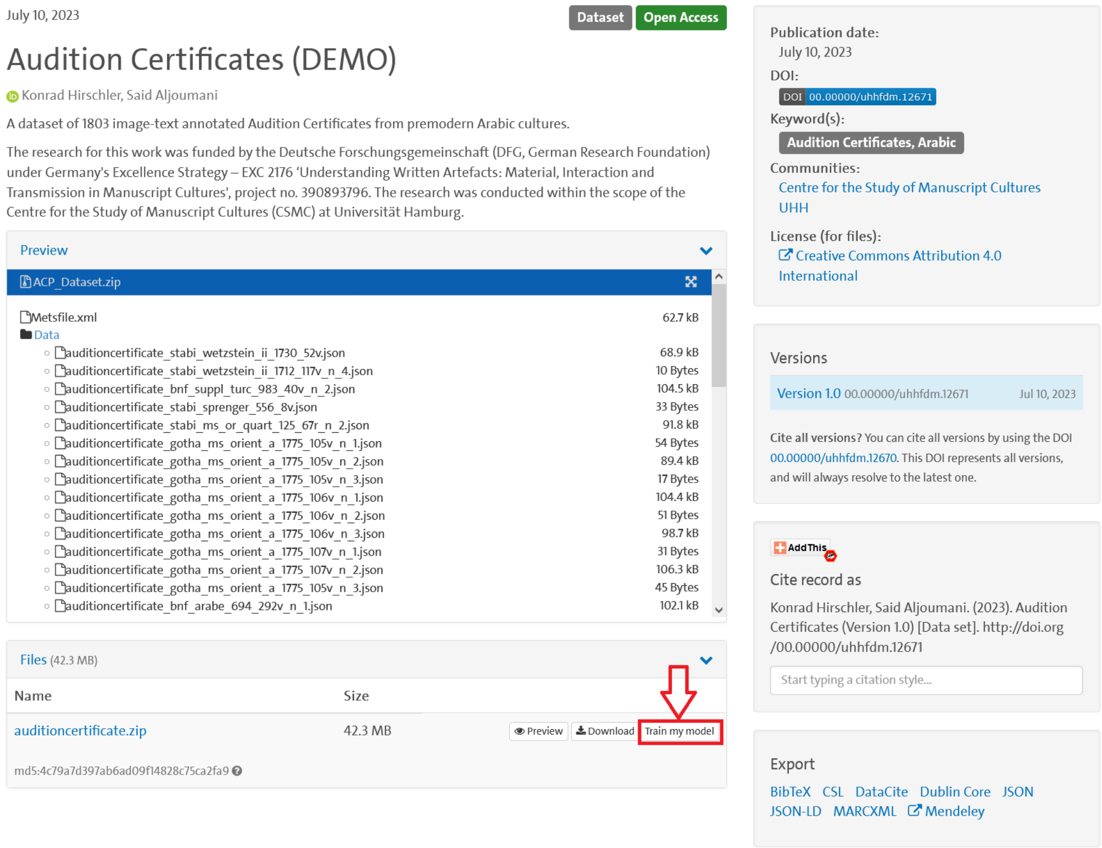


**Figure 2:** Research data repository with demo data. With red arrow and red border is the new "Train my model" button shown.

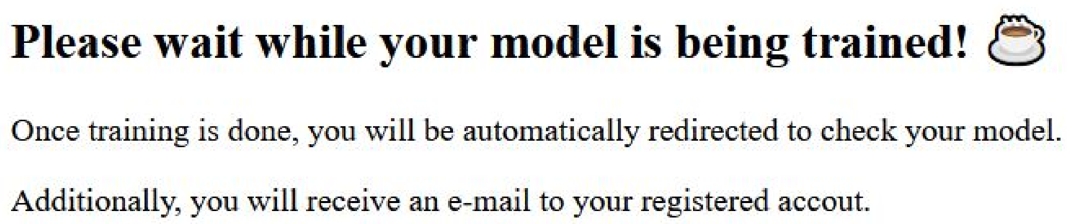


**Figure 3:** When the model is fine-tuned, a screen is shown to the user asking for patience.

In the screen, it is possible to enter a new text and see the predictions for two models. One is the pre-trained model, denoted as "Standard", while the other is the fine-tuned model denoted as "My model". This allows the user at a glance to decide whether they think that fine-tuning the model on the specified dataset has actually improved performance for their specific needs. In our example, one can see that, e.g. Islamic dates are now correctly identified which they were previously not.

If the user is satisfied with their model, they will be able to store it in the future back into the RDR. This will enable other users with a similar task, e.g. other Arabic texts that is not yet labelled but should be, to reuse the model for automatic labelling.

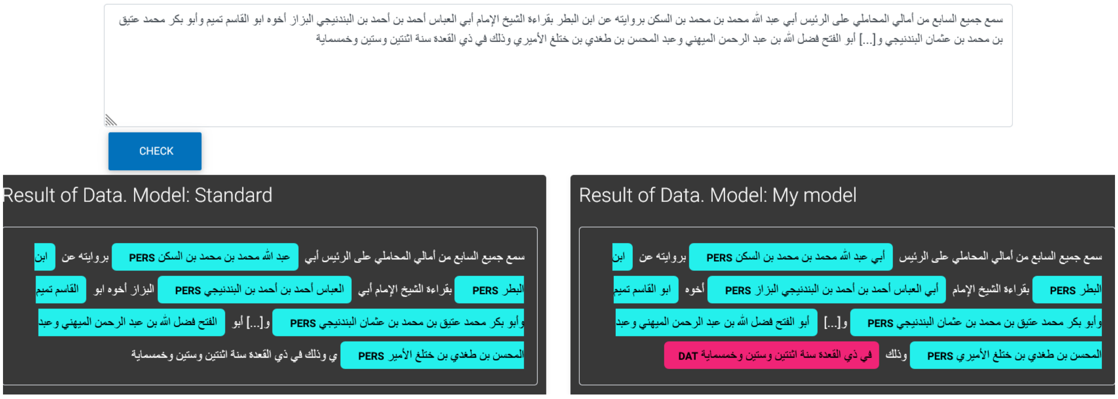


**Figure 4:** After fine-tuning is completed, this screen is shown to verify the model. “Standard” is in this case the pre-trained model, while “My model” is the fine-tuned. Text from MS al-Assad National Library Damascus, 3760/12, fol. 151r.

## 4. Application

Our demo application uses the audition certificates as the dataset for fine-tuning.

Fine-tuning was done using the HuggingFace Transformers library [13]. The pre-trained model of choice was “CAMeL-Lab/bert-base-arabic-camelbert-ca-ner”[6] which is a BERT model specifically pre-trained on classical Arabic texts and then fine-tuned on the ANERCorp dataset[7].

1800 of the 3000 annotated audition certificates in JSON-format have been loaded into a demo version of the research data repository in a Zip-file together with the METS-file. Annotations included persons, locations and dates.

Once everything has been uploaded correctly and the user is allowed to do fine-tuning with the specified data, the systems shows a new button called “Train my model” (Figure 2) which uses the parameters described in the METS-file. For our example, the following parameters have been chosen:

- A train-test-split of 80% and 20%,
- a learning rate of $1^{-4}$,
- a weight decay of $1^{-5}$, and
- 5 epochs.

All other parameters have been kept default as defined by the HuggingFace library. They can however also be changed, if necessary, by the user via METS-file.

Fine-tuning on an NVIDIA DGX2 with selected parameters took only a few minutes.

[6]https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-ca-ner
[7]https://camel.abudhabi.nyu.edu/anercorp/

## 5. Results and Evaluation

Performance of both the pre-trained CAMeLBERT as well as our fine-tuned model have been evaluated using the test dataset. We have evaluated the performance using precision, recall, and F1 score. Table 1 shows the results using the pre-trained model. It can be seen that three rows contain measures that are zero. The reason for the labels "MISC" as well as "ORG" having all zero scores is that both categories have not been labelled in our dataset. "DAT", i.e., date, was labelled 319 times but the pre-trained model was not able to detect them. Thus, they also got a zero score for all metrics. Locations as well as persons got some matches but the performance is not good enough using the test dataset.

| **label** | **precision** | **recall** | **F1 score** | **number** |
|---|---|---|---|---|
| DAT | 0.0 | 0.0 | 0.0 | 319 |
| LOC | 0.101 | 0.161 | 0.124 | 174 |
| MISC | 0.0 | 0.0 | 0.0 | 0 |
| ORG | 0.0 | 0.0 | 0.0 | 0 |
| PERS | 0.244 | 0.444 | 0.315 | 3115 |
| **Overall** | 0.225 | 0.391 | 0.285 | - |

**Table 1**
Performance of pre-trained BERT model

Table 2 provides the results using our fine-tuned model. Performance with all metrics has increased for all given labels. Since there are no entities labelled "MISC" and "ORG", both values have been omitted. As can be seen, performance has increased for all labels. The model is now able to detect dates which was not possible before. Especially looking at persons, the detection is now in a good range compared to the pre-trained model.

| **label** | **precision** | **recall** | **F1 score** | **number** |
|---|---|---|---|---|
| DAT | 0.464 | 0.571 | 0.512 | 319 |
| LOC | 0.534 | 0.586 | 0.559 | 174 |
| PERS | 0.798 | 0.853 | 0.825 | 3115 |
| **Overall** | **0.752** | **0.815** | **0.782** | - |

**Table 2**
Performance of fine-tuned BERT model

### 5.1. Problems and Challenges

Labelling is a task that is time consuming and monotonous. Thus, it can be prone to small errors and mistakes [14]. One mistake found in our dataset is the labelling of the و (Arabic translation of "and") as part of a person's name. This has an impact on both the fine-tuning of the model itself as well as on the computation of the metrics to evaluate the models.

# 6. ChatHA

Having employed a fine-tuning on demand system for identifying names of people, places, or dates, we aim to go further and outline a more feature rich and intuitive information system: Currently, the model fine-tuned on demand is made available to an information system providing a graphical user interface. However, it is sometimes difficult to fiddle with graphical user interfaces and each interface needs to be created for the specific tasks of humanities scholars. Additionally, the scholars need to get to know each new interface. Instead, we would like to build a system with a more general interface, allowing the scholars to interact in a more natural way with the data in the model fine-tuned on demand.

As solution, we outline ChatHA, a Humanities Aligned Chatbot. We adopt the already presented mechanism to fine-tune an LLM, e.g., GPT-4 [15]. The fine-tuned model can be used to provide a chatbot that can answer natural language questions about ancient texts and the humanities. In contrast to publicly available chatbots, ChatHA gets fine-tuned on the specific data in the research data repository. Hence, ChatHA is able to provide detailed answers based on the available data. Additionally, the ability to process natural language questions and providing the answers the same way lowers the barriers for the humanities scholars to use the system. It is not necessary to create a graphical user interface for a task, as each task can be send as textual question to ChatHA.

However, a system like ChatHA built automatically and with less supervisions also implies some issues: LLMs have no *true* understanding of the research data repository and its content, they try to combine the best answer based on texts they processed during training. Therefore, LLMs are prone to hallucinations, invented answers, and do not cite their sources. To combat this, we do not output the raw LLM output during question answering, but rather post-process it to include citations. The obtained result is then displayed to the user and the citations allow the user to validate the answer.

We now describe this post-processing and the overall workflow in more detail: First, we choose some pre-trained LLM. We opt for using a pre-trained version to include basic natural language understanding and general query answering. Second, the user, in our case a humanities scholar, chooses on which types of texts the LLM should be fine-tuned on. The texts can be, e.g., Arabic or Tamil[8] or the whole research data repository. This step composes the corpus for our LLM and ensures the alignment for humanities of the fine-tuned LLM. Third, we fine-tune the LLM with the selected data and create the chatbot by this step. However, we still have the issue of hallucinations and missing citations.

As a solution, we apply Subjective Content Descriptions (SCDs) [16]. SCDs are additional data attachted to locations in text documents, i.e., an SCD contains additional data like desciptions, links, or labels which themselves may be created automatically or by humans and each SCD is attachted to one or more sentences of a text document. Additionally, the sentences an SCD is attached to are represented in an SCD word-distribution matrix. Using this SCD word-distribution matrix, an SCD can be identified by the Most Probably Suited SCD (MPS$^2$CD) algorithm [17] for any new and unseen sentence. MPS$^2$CD identifies the most suitable SCD from the set of known SCDs attached to the text documents. Hence, using MPS$^2$CD it is possible to

[8]https://www.awhamburg.de/forschung/langzeitvorhaben/tamilex.html

create a link from a new and unseen sentence to an SCD and all sentences this SCD is attached with. Generally, the theory of SCDs is not restricted to text documents attachted with SCD containing additional text.

Coming back to ChatHA, we lack SCDs on the corpus used for fine-tuning. The UnSupervised Estimator of SCD Matrices (USEM) [18] automatically creates SCDs for a corpus and the SCDs get attached to the sentences in the corpus, too. Thus, using USEM we add SCDs to the corpus used for fine-tuning. Each SCD represents a topic or concept mentioned in the corpus and all sentences about each topic or concept belong to the same SCD. Thus, our SCDs represents the various topics or concepts in the corpus and the sentences that mention them.

Using SCDs we can solve the issue of hallucinations and missing citations in the output of the LLM. For each sentences in the output, MPS$^2$CD identifies an SCD from the corpus. In doing so, a link from the output of the LLM to the SCDs of the corpus and further on to sentences of the corpus is created. These links can now be used as citations shown in the output pointing to relevant sentence in the corpus used for fine-tuning.

Finally, ChatHA is ready to be used: A humanities scholar inputs a question about Arabic or Tamil texts using natural language. This question is first sent to the LLM and its outputs is post-processed in the following way: We apply MPS$^2$CD on the raw output to identify an SCD for each sentence. Sentences for which no SCD is found may be hallucinations and are omitted because there is no evidence of SCDs. The processed output is then displayed to the user alongside with the SCDs for each sentence. For each sentence and SCD, ChatHA offers the possibilty to view this SCD in the information system for USEM [19] or to open each of the sentences in the corpus used for fine-tuning which are attached to the same SCD. Additionally, if further visualization is available for an SCD or a sentence, ChatHA offers to visualize it in the corresponding information system, e.g., for Arabic texts the system shown in Figure 4.

All in all, ChatHA can be used to query the research data repository for research tasks in the humanities. The output includes citations, so we not only reduce hallucinations, but also give pointers for a more detailed look.

However, some questions might go beyond the information available in the corpus and the research data repository. For this case, it should be possible to include corresponding academic publications and further web resources about humanities in the corpus. Such resources might be obtained by crawling the web and asking multiple questions to ChatHA, maybe also combined with running a second fine-tuning on the new corpus. AutoGPT [20] is a chatbot for more sophisticated tasks which may require multiple questions. It will split a task in multiple questions, run each question and the implied tasks, before it combines the answers with some automated answer combination mechanism. Thus, AutoGPT can be used to answer more challenging questions that require deeper understanding and expressiveness.

## 7. Conclusion

This paper introduces the FToD process, a process that helps people that are not experts in AI training a custom NLP model. Data already available in a research data repository including a METS-file can be used directly without knowing the specifics of the libraries used. It was shown, using the ACP training data and CAMeLBERT, that the process is able to improve results using labelled data already available in the repository.

After fine-tuning the model, the expert is able to evaluate the model. As a next step, it is planned to use that for the expert to enter new samples into the research data management. A new text is entered and the results are shown. It should then be possible to change the labels when needed and to pick labels from the pre-trained model if they are seen as better. Afterwards, this new example should be part of the dataset which can be used to fine-tune a potentially better model.

Future work focuses also on the implementation and refinement of ChatHA.

## Acknowledgments

The research for this contribution was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2176 'Understanding Written Artefacts: Material, Interaction and Transmission in Manuscript Cultures', project no. 390893796.

## References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[2] C. Lei, Unsupervised Learning: Word Vector, Springer Singapore, Singapore, 2021, pp. 95–149. URL: https://doi.org/10.1007/978-981-16-2233-5_7. doi:10.1007/978-981-16-2233-5_7.

[3] D. Rothman, Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more, Packt Publishing, 2021. URL: https://books.google.de/books?id=Cr0YEAAAQBAJ.

[4] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[5] S. Kublik, S. Saboo, GPT-3: The Ultimate Guide To Building NLP Products With OpenAI API, Packt Publishing, 2023. URL: https://books.google.de/books?id=fgutEAAAQBAJ.

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, CoRR abs/1910.10683 (2019). URL: http://arxiv.org/abs/1910.10683. arXiv:1910.10683.

[7] E. Webber, A. Olgiati, Pretrain Vision and Large Language Models in Python: End-to-end techniques for building and deploying foundation models on AWS (English Edition), 1 ed., Packt Publishing; 1. Edition (31. Mai 2023), 2023, pp. 133–219.

[8] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, N. Habash, The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models, in: Proceedings of the Sixth Arabic

Natural Language Processing Workshop, Association for Computational Linguistics, Kyiv, Ukraine (Online), 2021.
[9] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heinz, D. Roth, Recent advances in natural language processing via large pre-trained language models: A survey, 2021. arXiv:2111.01243.
[10] C. Tilp, Compilation and preview of archieved research data with a manifest-file, Bachelor thesis, Universität zu Lübeck, 2023.
[11] S. Schiff, S. Melzer, E. Wilden, R. Möller, TEI-Based Interactive Critical Editions, in: S. Uchida, E. Barney, V. Eglin (Eds.), Document Analysis Systems, Springer International Publishing, Cham, 2022, pp. 230–244.
[12] N. Stahl, New concepts for previewing of data management repositories, Bachelor thesis, Universität zu Lübeck, 2023.
[13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.
[14] C. G. Northcutt, A. Athalye, J. Mueller, Pervasive label errors in test sets destabilize machine learning benchmarks, 2021. arXiv:2103.14749.
[15] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.
[16] F. Kuhr, T. Braun, M. Bender, R. Möller, To Extend or not to Extend? Context-specific Corpus Enrichment, Proceedings of AI: Advances in Artificial Intelligence (2019) 357–368. doi:10.1007/978-3-030-35288-2_29.
[17] F. Kuhr, M. Bender, T. Braun, R. Möller, Augmenting and automating corpus enrichment, Int. J. Semantic Computing 14 (2020) 173–197. doi:10.1142/S1793351X20400061.
[18] M. Bender, T. Braun, R. Möller, M. Gehrke, Unsupervised estimation of subjective content descriptions, Proceedings of the 17th IEEE International Conference on Semantic Computing (ICSC-23) (2023). doi:10.1109/ICSC56153.2023.00052.
[19] M. Bender, T. Braun, R. Möller, M. Gehrke, Unsupervised estimation of subjective content descriptions in an information system, Int. Journal of Semantic Computing (2023).
[20] H. Yang, S. Yue, Y. He, Auto-gpt for online decision making: Benchmarks and additional opinions, 2023. arXiv:2306.02224.