

AI-Based Deep Indexing of Historical Mass Files

Gernot A. Fink

TU Dortmund, Department of Computer Science

Lamarr-Institute for Machine Learning and Artificial Intelligence

Universität Münster, February 27, 2026

With support of:

LAV NRW (sample archival materials), T. Hallyburton, A. Matei, O. Tüselmann, F. Wolf

Introduction

If you have a mass(!) indexing problem ...



Image source: Census 1950, LAV NRW; Gernot A. Fink

Manual doesn't labor (really) help anymore!

⇒ Automation absolutely essential!

Introduction

If you have a mass(!) indexing problem ...



Image source: Census 1950, LAV NRW; Gernot A. Fink

Manual doesn't labor (really) help anymore!

⇒ Automation absolutely essential!

Introduction

If you have a mass(!) indexing problem ...



Image source: Census 1950, LAV NRW; Gernot A. Fink

Manual doesn't labor (really) help anymore!

⇒ Automation absolutely essential!

Introduction

If you have a mass(!) indexing problem ...



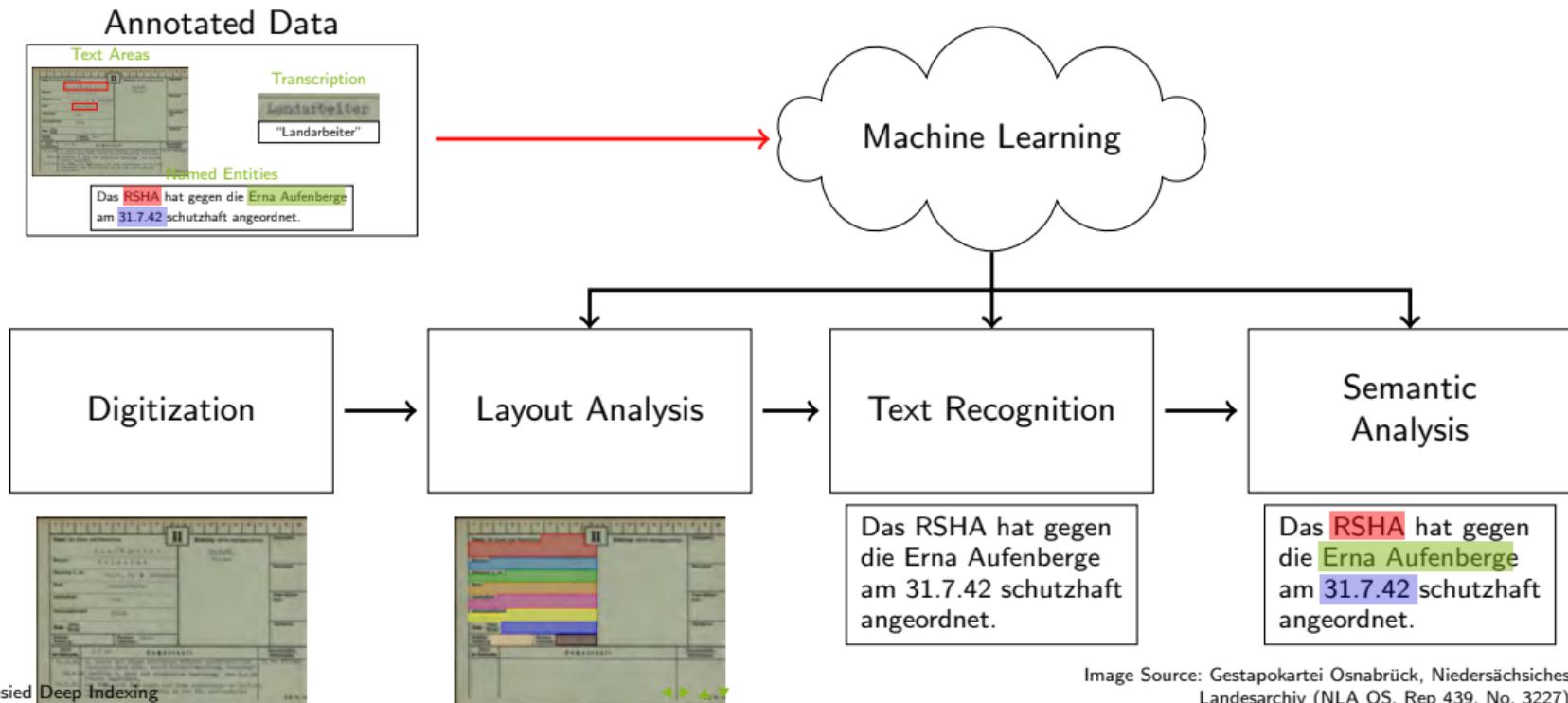
Image source: Census 1950, LAV NRW; Gernot A. Fink

Manual doesn't labor (really) help anymore!

⇒ Automation absolutely essential!

Document Analysis with Future Prospects

Document Analysis with Machine Learning



Document Analysis with Generative AI

- ▶ Advances in generative AI models (ChatGPT and “friends”)
- ▶ Model Type: Large Vision Language Models
- ▶ Integrated analysis in *a single model/formalism!*

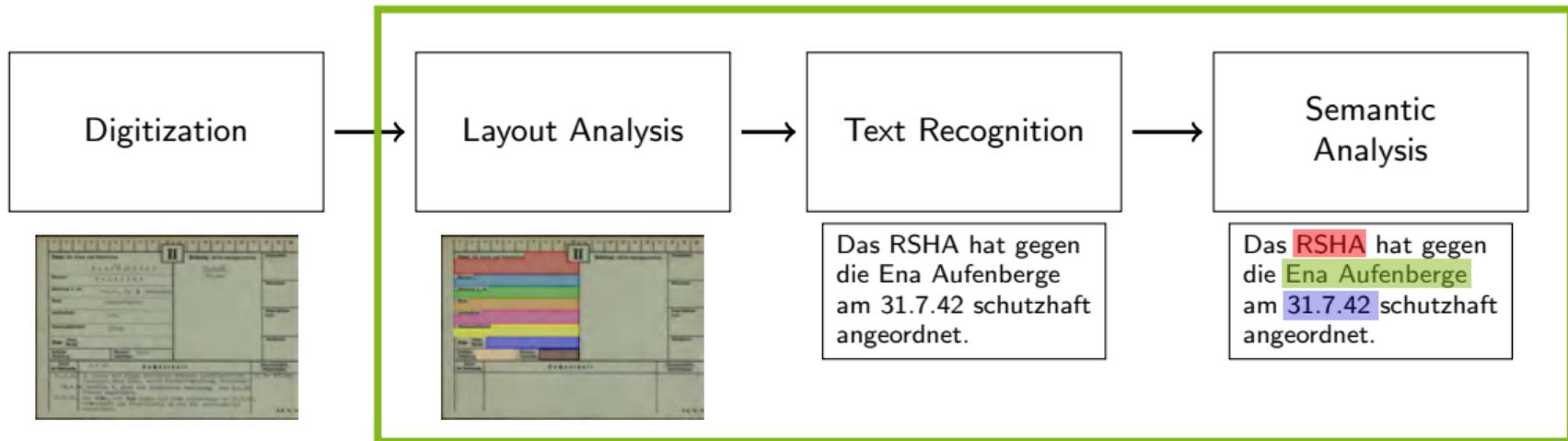
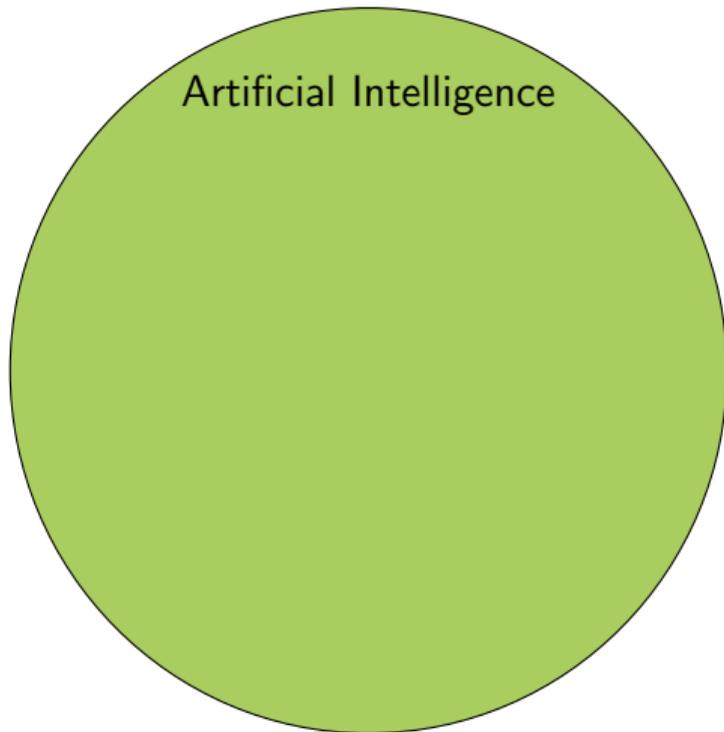
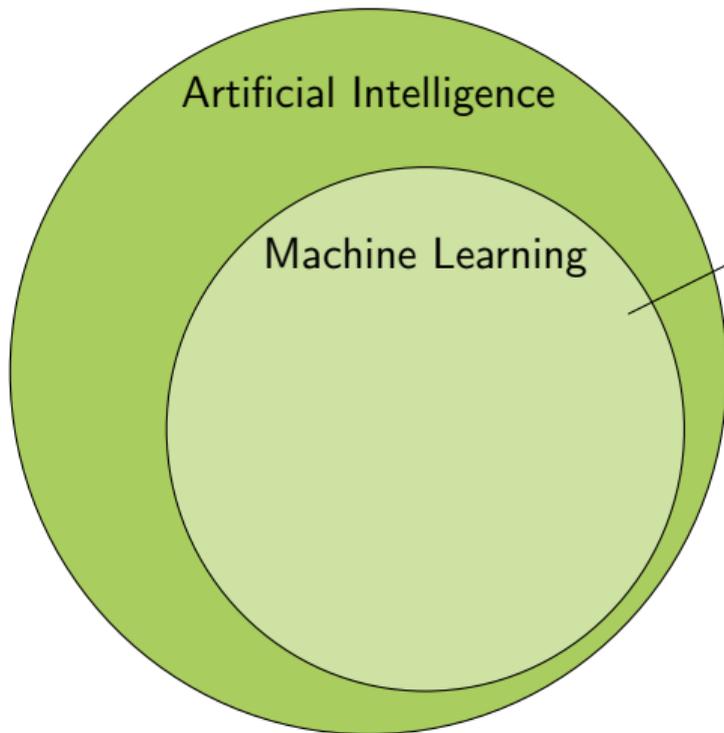


Image source: Gestapokartei Osnabrück, Niedersächsisches Landesarchiv (NLA OS, Rep 439, No. 3227)

Artificial Intelligence: Terminology



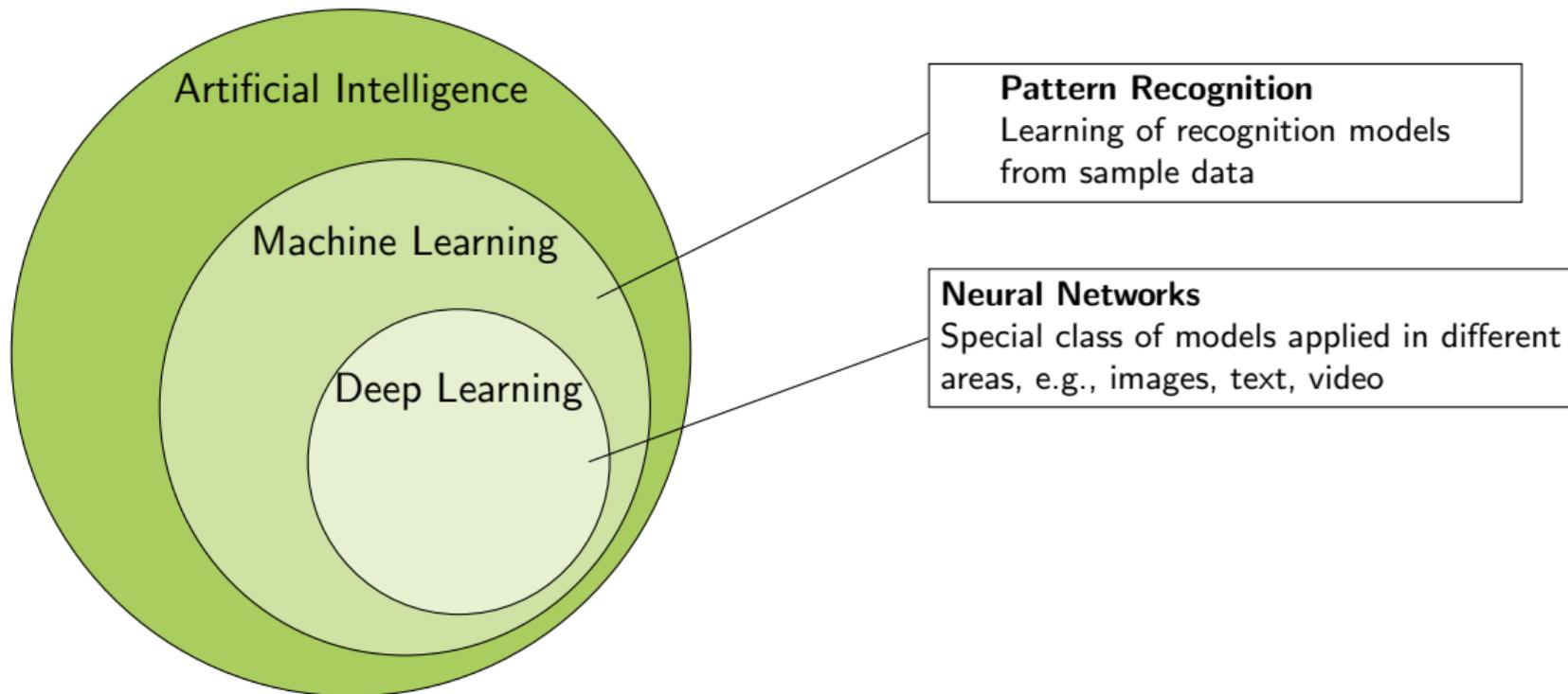
Artificial Intelligence: Terminology



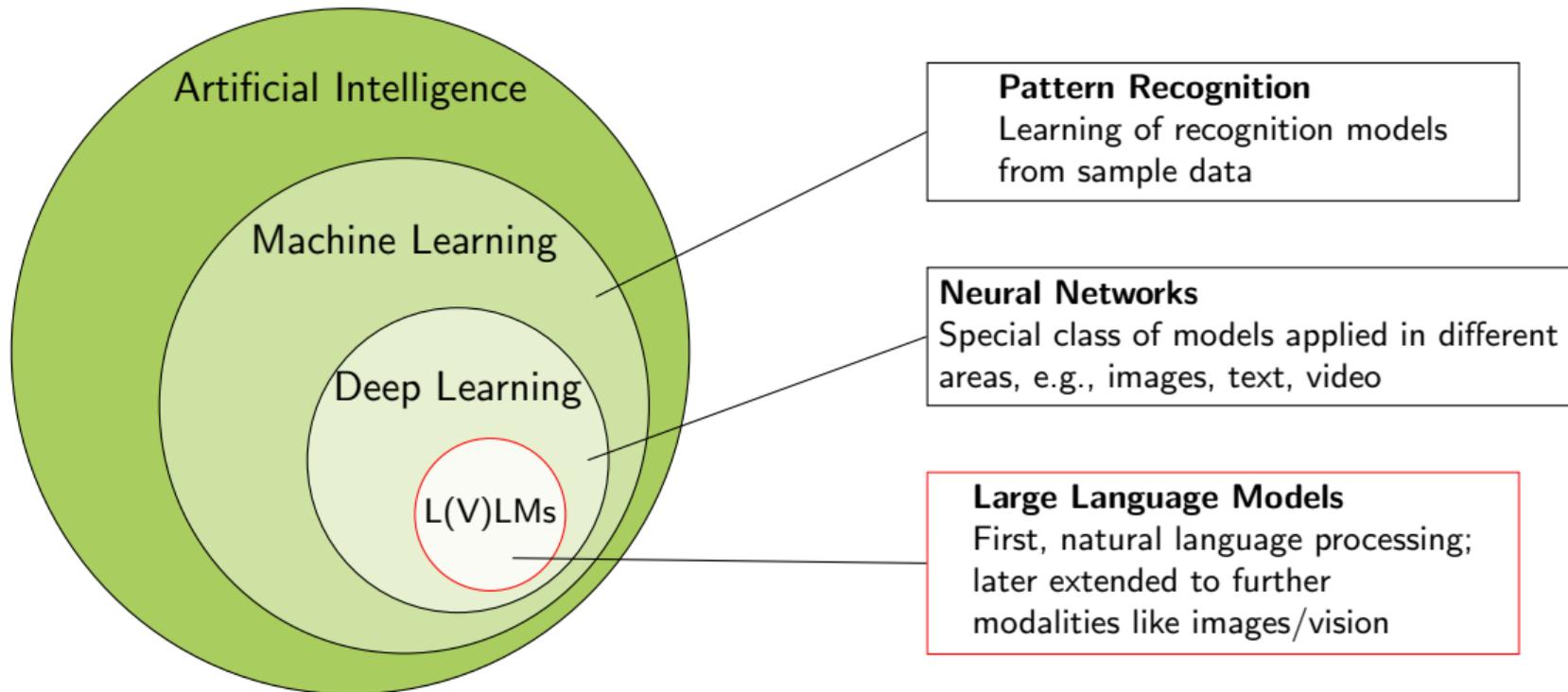
Pattern Recognition

Learning of recognition models from sample data

Artificial Intelligence: Terminology

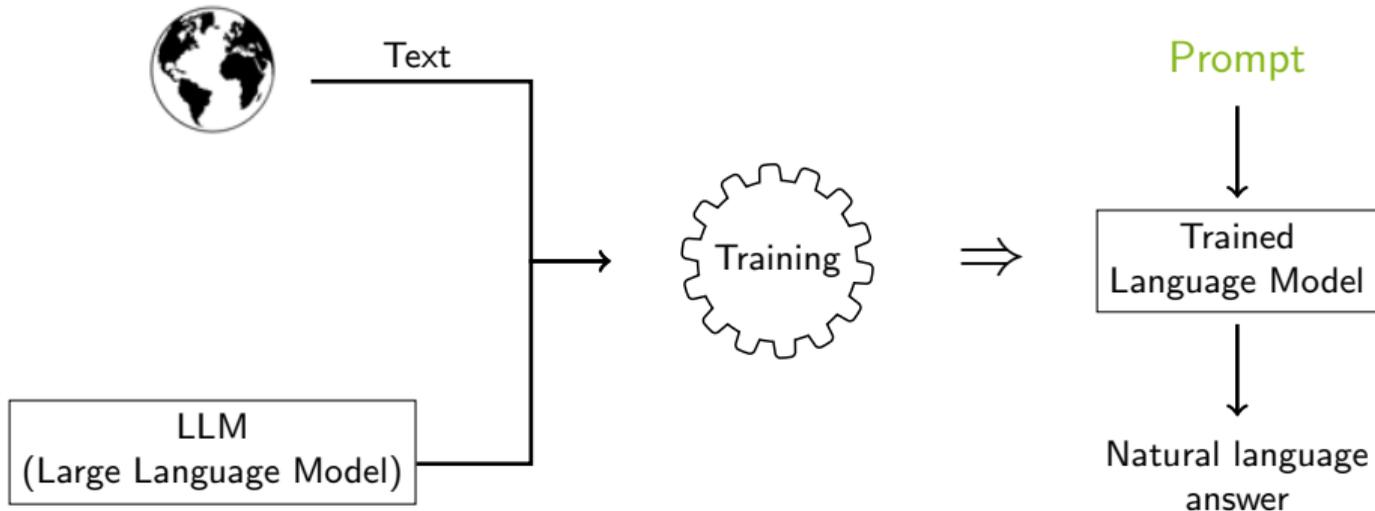


Artificial Intelligence: Terminology



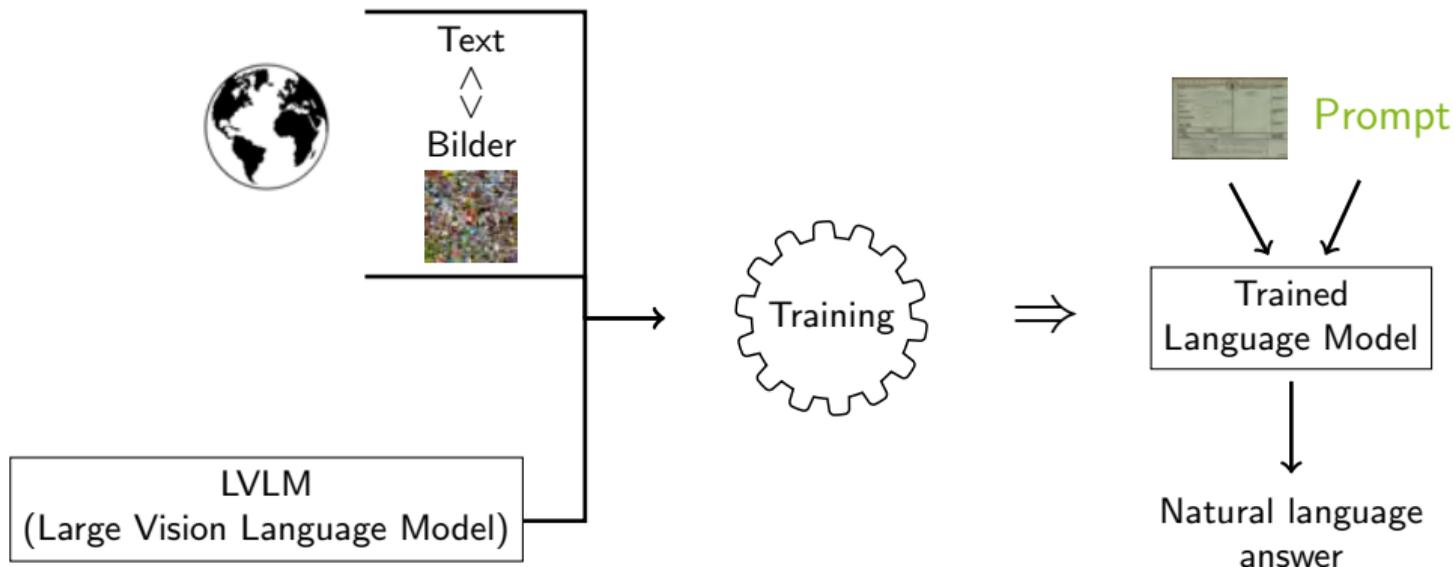
Large Language Models

- ▶ No problem defined/given beforehand!
- ▶ **Goal:** Generation of natural language
- ▶ Training requires no specially annotated training data



Large Vision Language Models

- ▶ Extension to process images (= vision) as input
- ▶ Allows the generation of language output based on an input image



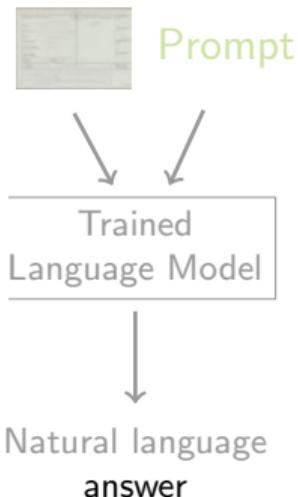
Large Vision Language Models

- ▶ Extension to process images (= vision) as input
- ▶ Allows the generation of language output based on an input image



- ▶ Training data from the Internet
- ▶ Well suited for problems covered
- ▶ Creation/operation *extremely* expensive!

LVLM
(Large Vision Language Model)



Example: Census of England and Wales 1911

CENSUS OF ENGLAND AND WALES, 1911.

Before writing on this Schedule please read the Examples and the Instructions given on the other side of the paper, as well as the headings of the Columns. The entries should be written in ink.

The contents of the Schedule will be treated as confidential. Great care will be taken that no information is disclosed with regard to individual persons. The returns are not to be used for proof of age, or in connection with Old Age Pensions, or for any other purpose than the preparation of Statistical Tables.

NAME AND SURNAME	RELATIONSHIP TO HEAD OF FAMILY	AGE AND SEX	PARTICULARS AS TO MARRIAGE				PROFESSION OR OCCUPATION OF PERSON aged 16 years and upwards.				NATIONALITY OF BORN ABROAD	INFIRMITY	
			Married	Single	Widow	Divorced	Employed	Unemployed	Retired	Other			
1. Geo. H. Hayward	Head	47	Married	12	9	3	Labourer	571	Can do manual	—	—	—	—
2. Annie Hayward	Wife	47	Married	24	—	—	at Home	—	—	—	—	—	—
3. George H. Hayward	Son	20	Single	—	—	—	Art. Smith	Widley 79	Widley	—	—	—	—
4. Joseph Hayward	Son	18	Single	—	—	—	Printer	224	Widley	—	—	—	—
5. William J. Hayward	Son	14	Single	—	—	—	Shedman	Widley 90	Widley	—	—	—	—
6. Henry Hayward	Son	14	Single	—	—	—	Grand	Widley 330	Widley	—	—	—	—
7. Mary J. Hayward	Daughter	13	Single	—	—	—	School	390	—	—	—	—	—
8. Alfred H. Hayward	Son	8	Single	—	—	—	School	—	—	—	—	—	—
9. Thomas Hayward	Son	5	Single	—	—	—	School	—	—	—	—	—	—

(To be filled up by the Enumerator.)

Total

24

42

9

(To be filled up by, or on behalf of, the Head of Family or other person in occupation, or in charge, of the dwelling.)

I declare that this Schedule is correctly filled up to the best of my knowledge and belief.

Signature: *George H. Hayward*

Printed Address: *Widley, Wilt. England*

¹Image source: Kresen Kernov ('Cornwall Centre'), UK

Example: Census of England and Wales 1911

NAME AND SURNAME	RELATIONSHIP to Head of Family.	AGE (last Birthday) and SEX.
of every Person, whether Member of Family, Visitor, Boarder, or Servant, who		For Infants under one year state the age in months or "under one month," "one month," etc.
(1) passed the night of Sunday, April 2nd, 1911, in this dwelling and was alive at midnight, or	State whether "Head," or "Wife," "Son," "Daughter," or other relative, "Visitor," or "Servant."	
(2) arrived in this dwelling on the morning of Monday, April 3rd, not having been enumerated elsewhere.		
No one else must be included. (For order of entering names see Examples on back of Schedule.)		Age of Males. Age of Females.
1.	2.	3.
1 Geo. H. Kelynack	Head	47 -
2 Janie Kelynack	Wife	- 47
3 George H. Kelynack	Son	20 -
4 Joseph Kelynack	Son	18 -
5 William J. Kelynack	Son	16 -
6 Henry Kelynack	Son	14 -
7 Mary A. Kelynack	Daughter	13
8 Alfred H. Kelynack	Son	8 -
9 Thomas Kelynack	Son	5 -
10		

- ▶ **Prompt:** What is the name of the head of family?
- ▶ The head of family is "Geo H. Kelynack".
- ▶ **Prompt:** How many persons are there?
- ▶ In this document there are in total 9 persons.
- ▶ **Prompt:** How many sons does Geo Kelynack have?
- ▶ Geo H. Kelynack as 6 sons in total.
- ▶ **Prompt:** How many persons are female?
- ▶ According to the document there are two female persons.

¹Image source: Kresen Kernov ('Cornwall Centre'), UK

²Results generated with OpenAI, GPT4, prompted in German

Example: Census of England and Wales 1911

- **Prompt:** Provide me with the information from the document as a structured file (JSON).

NAME AND SURNAME	RELATIONSHIP to Head of Family.	AGE (last Birthday) and SEX.	
		Age of Males.	Age of Females.
of every Person, whether Member of Family, Visitor, Boarder, or Servant, who (1) passed the night of Sunday, April 2nd, 1911, in this dwelling and was alive at midnight, or (2) arrived in this dwelling on the morning of Monday, April 3rd, not having been enumerated elsewhere. No one else must be included. (For order of entering names see Examples on back of Schedule.)	State whether "Head," or "Wife," "Son," "Daughter," or other Relative, "Visitor," "Boarder," or "Servant."	For Infants under one year state the age in months or "under one month," "one month," etc.	
		1	2
1 Geo. H. Kelynack	Head	47	—
2 Janie Kelynack	Wife	—	47
3 George H. Kelynack	Son	20	—
4 Joseph Kelynack	Son	18	—
5 William J. Kelynack	Son	16	—
6 Henry Kelynack	Son	14	—
7 Mary A. Kelynack	Daughter	13	—
8 Alfred H. Kelynack	Son	8	—
9 Thomas Kelynack	Son	5	—

{
{
,
{
,
,
...
}

Name and Surname: Geo H. Kelynack,
Relationship to Head of Family: Head,
Age (last Birthday) and Sex: 47 -

Name and Surname: Janie Kelynack,
Relationship to Head of Family: Wife,
Age (last Birthday) and Sex: - 47

¹Image source: Kresen Kernow ('Cornwall Centre'), UK
²Results generated with OpenAI, GPT4, prompted in German

Example: Census of England and Wales 1911

- **Prompt:** Provide me with the information from the document as a structured file (JSON).

NAME AND SURNAME	RELATIONSHIP to Head of Family.	AGE (last Birthday) and SEX.	
		Age of Males.	Age of Females.
of every Person, whether Member of Family, Visitor, Boarder, or Servant, who (1) passed the night of Sunday, April 2nd, 1911, in this dwelling and was alive at midnight, or (2) arrived in this dwelling on the morning of Monday, April 3rd, not having been enumerated elsewhere. No one else must be included. (For order of entering names see Examples on back of Schedule.)	State whether "Head," or "Wife," "Son," "Daughter," or other Relative, "Visitor," "Boarder," or "Servant."	For Infants under one year state the age in months or "under one month," "one month," etc.	
1	Geo H Kelynack	Head	47 -
2	Janie Kelynack	Wife	- 47
3	George H Kelynack	Son	20 -
4	Joseph Kelynack	Son	18 -
5	William J Kelynack	Son	16 -
6	Henry Kelynack	Son	14 -
7	Mary A Kelynack	Daughter	13 -
8	Alfred H Kelynack	Son	8 -
9	Thomas Kelynack	Son	5 -
10			

{

{

Name and Surname: Geo H. Kelynack,
Relationship to Head of Family: Head,
Age (last Birthday) and Sex: 47 -

},

{

Name and Surname: Janie Kelynack,
Relationship to Head of Family: Wife,
Age (last Birthday) and Sex: - 47

},

...

}

Impressive results, but ...

¹Image source: Kresen Kernov ('Cornwall Centre'), UK

²Results generated with OpenAI, GPT4, prompted in German

Example: Census of England and Wales 1911

- **Prompt:** Provide me with the information from the document as a structured file (JSON).

NAME AND SURNAME	RELATIONSHIP to Head of Family.	AGE (last Birthday) and SEX.	
		Age of Males.	Age of Females.
of every Person, whether Member of Family, Visitor, Boarder, or Servant, who (1) passed the night of Sunday, April 2nd, 1911, in this dwelling and was alive at midnight, or (2) arrived in this dwelling on the morning of Monday, April 3rd, not having been enumerated elsewhere. No one else must be included. (For order of entering names see Examples on back of Schedule.)	State whether "Head," or "Wife," "Son," "Daughter," or other Relative, "Visitor," "Boarder," or "Servant."	For Infants under one year state the age in months or "under one month," "one month," etc.	
1	Geo H Kelynack	Head	47 -
2	Janie Kelynack	Wife	- 47
3	George H Kelynack	Son	20 -
4	Joseph Kelynack	Son	18 -
5	William J Kelynack	Son	16 -
6	Henry Kelynack	Son	14 -
7	Mary A Kelynack	Daughter	13
8	Alfred H Kelynack	Son	8 -
9	Thomas Kelynack	Son	5 -
10			

{

{

Name and Surname: Geo H. Kelynack,
Relationship to Head of Family: Head,
Age (last Birthday) and Sex: 47 -

},

{

Name and Surname: Janie Kelynack,
Relationship to Head of Family: Wife,
Age (last Birthday) and Sex: - 47

},

}

... **Problem #1: Data protection**

¹Image source: Kresen Kernov ('Cornwall Centre'), UK

²Results generated with OpenAI, GPT4, prompted in German

Example: Census of England and Wales 1911

- **Prompt:** Provide me with the information from the document as a structured file (JSON).

NAME AND SURNAME	RELATIONSHIP to Head of Family.	AGE (last Birthday) and SEX.	
		Age of Males.	Age of Females.
of every Person, whether Member of Family, Visitor, Boarder, or Servant, who (1) passed the night of Sunday, April 2nd, 1911, in this dwelling and was alive at midnight, or (2) arrived in this dwelling on the morning of Monday, April 3rd, not having been enumerated elsewhere. No one else must be included. (For order of entering names see Examples on back of Schedule.)	State whether "Head," or "Wife," "Son," "Daughter," or other Relative, "Visitor," "Boarder," or "Servant."	For Infants under one year state the age in months or "under one month," "one month," etc.	
1	Geo H Kelynack	Head	47 -
2	Janie Kelynack	Wife	- 47
3	George H Kelynack	Son	20 -
4	Joseph Kelynack	Son	18 -
5	William J Kelynack	Son	16 -
6	Henry Kelynack	Son	14 -
7	Mary A Kelynack	Daughter	13 -
8	Alfred H Kelynack	Son	8 -
9	Thomas Kelynack	Son	5 -
10			

{

{

Name and Surname: Geo H. Kelynack,
Relationship to Head of Family: Head,
Age (last Birthday) and Sex: 47 -

},

{

Name and Surname: Janie Kelynack,
Relationship to Head of Family: Wife,
Age (last Birthday) and Sex: - 47

},

}

... **Solution: Open Source Model**

¹Image source: Kresen Kernov ('Cornwall Centre'), UK

²Results generated with OpenAI, GPT4, prompted in German

Example: Field Postcard (dt. "Feldpostkarte")



- ▶ **Prompt:** What is the name of the recipient?
- ▶ The name of the recipient is Frau M. Lubor, below which is written Hr. R. Schwan.
- ▶ **Prompt:** What is the destination address of the postcard?
- ▶ Fulda, Mehlerstraße 17 II
- ▶ **Prompt:** Which date and which location are given on the stamp?
- ▶ The date is 12.4.17 and the location Zweibrücken.

¹Image source: Bley, B: Feldpostkarten im 1. Weltkrieg, Private Collection

²Results generated with OpenAI, GPT4, prompted in German

Example: Field Postcard (dt. "Feldpostkarte")



- ▶ **Prompt:** What is the name of the recipient?
- ▶ The name of the recipient is Frau M. Lubor, below which is written Hr. R. Schwan.

- ▶ **Prompt:** What is the destination address of the postcard?
- ▶ Fulda, Mehlerstraße 17 II

- ▶ **Prompt:** Which date and which location are given on the stamp?
- ▶ The date is 12.4.17 and the location Zweibrücken.

¹Image source: Bley, B: Feldpostkarten im 1. Weltkrieg, Private Collection

²Results generated with OpenAI, GPT4, prompted in German

Example: Compensation Files (WW II)

Antrag
auf Grund des Bundesergänzungsgesetzes zur Entschädigung für Opfer der nationalsozialistischen Verfolgung (BEG)
vom 18. 9. 1953 (BGBl. I S. 1387)

I. Anspruchsberechtigte(r)

1. Name: Bruno
 Vornamen: Matthias
 Geburtstag und -ort (Kreis, Land): 10.10.1884 in Bölsdorf
des Bismarck-Kr. NRW.
 Jetziger Wohnort / Sitz (Kreis, Land): Bismarck-Kreis, Bismarckstr. 217
 (Straße und Haus-Nr.)
 Familienstand: led. / verh. / verw. / gesch.
 Anzahl der Kinder: zwei Alter der Kinder: 36 + 35 Jahre
 Staatsangehörigkeit: frühere: deutsch jetzige: deutsch
 2. Beruf: Gast- u. Handwirt
 Erlernter Beruf: Restaurateur

Prompt: What is the name of the authorized person?

Prompt: What is the given name of the authorized person?

Prompt: What is the authorized person's profession?

Walter

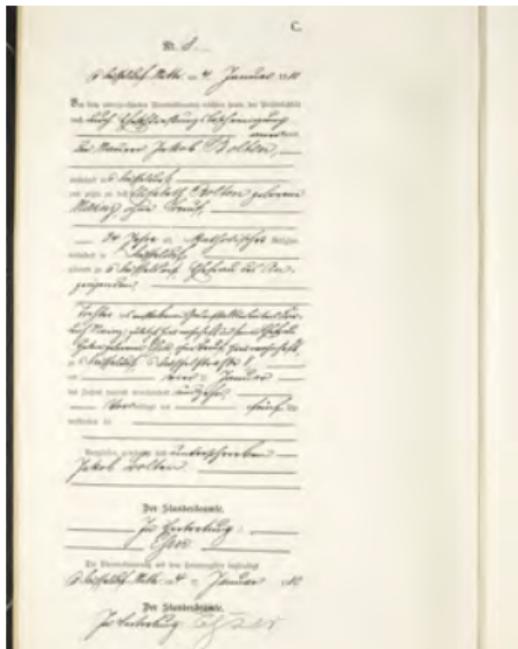
Walter

Gastwirt in Hotelien

¹Image source: LAV NRW R, BR 3007 Nr. 699

²Results generated with Google, Paligemma 3B-ft-docvqa-448, prompted in German

Example: Civil Registry Documents



- ▶ Prompt: Who appeared before the registrar today?
- ▶ mr. schultze
- ▶ Prompt: On what date was the document written?
- ▶ 4. November 1910

¹Image source: LAV NRW R, PA 3103 Nr. 4750

²Results generated with Google, Paligemma 3B-ft-docvqa-448, prompted in German

Example: Civil Registry Documents

Vor dem unterzeichneten Standesbeamten erschien heute, der Persönlichkeit
nach Herrn Schulze Wolfgang Schulze
_____ mann,
Herrn Mannen Johann Polke, —

am 4. Januar 1910

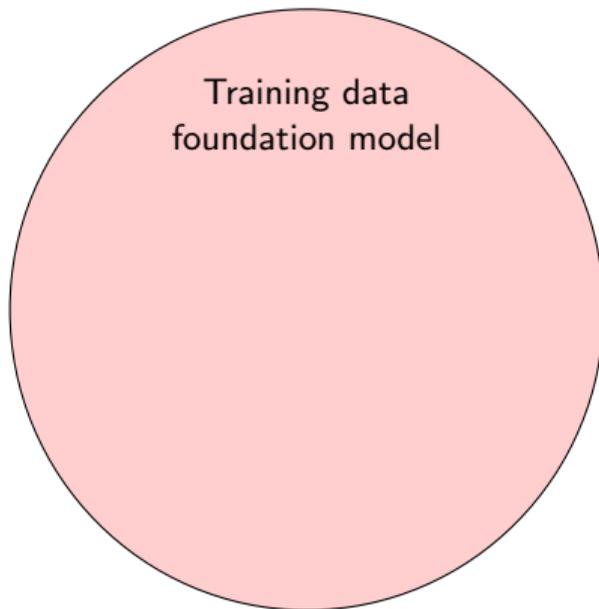
- ▶ **Prompt:** Who appeared before the registrar today?
- ▶ **mr. schultze**

- ▶ **Prompt:** On what date was the document written?
- ▶ **4. November** 1910

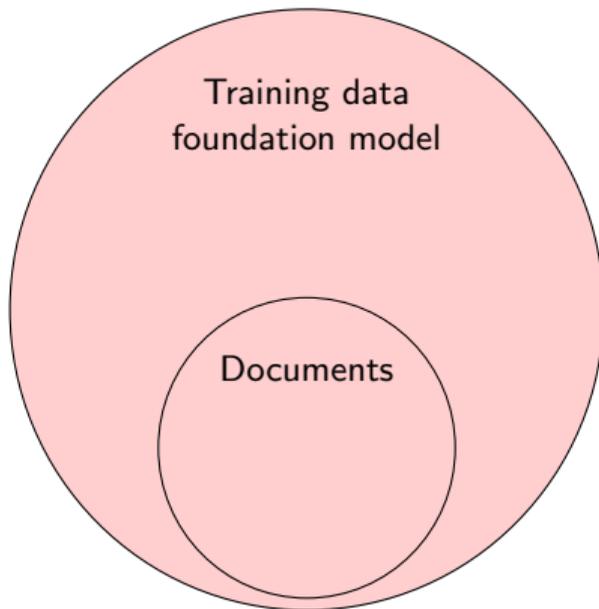
¹Image source: LAV NRW R, PA 3103 Nr. 4750

²Results generated with Google, Paligemma 3B-ft-docvqa-448, prompted in German

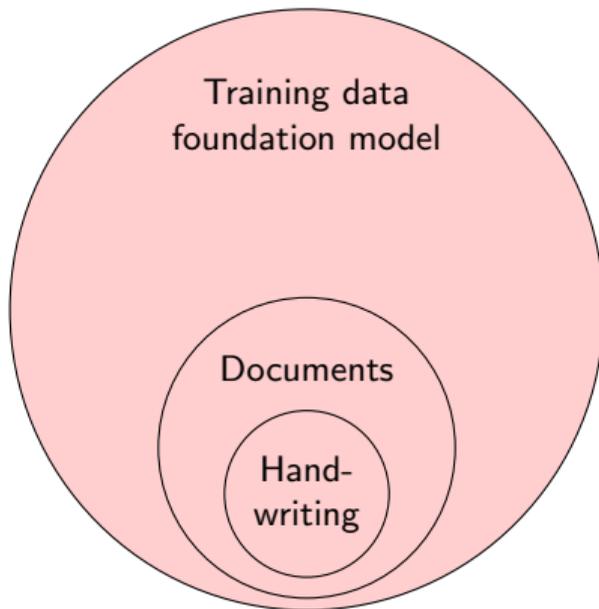
Problem #2: Insufficient Domain Coverage!



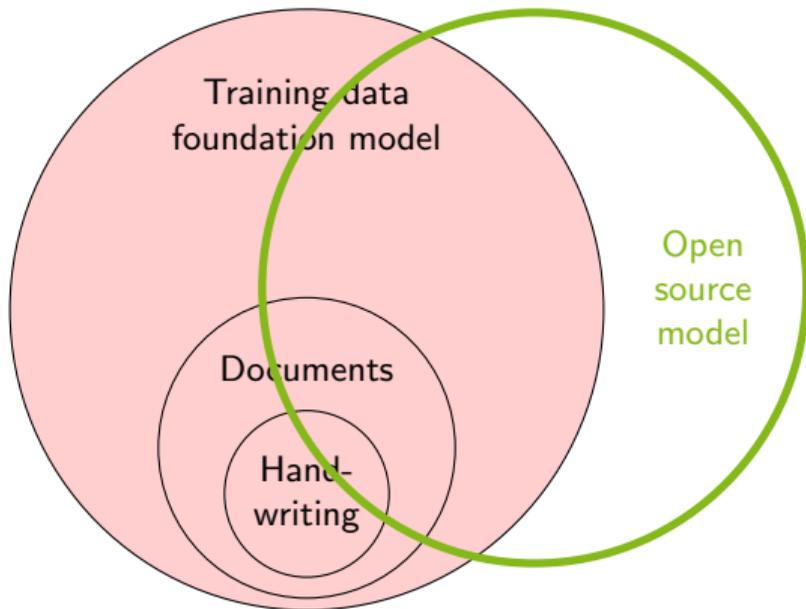
Problem #2: Insufficient Domain Coverage!



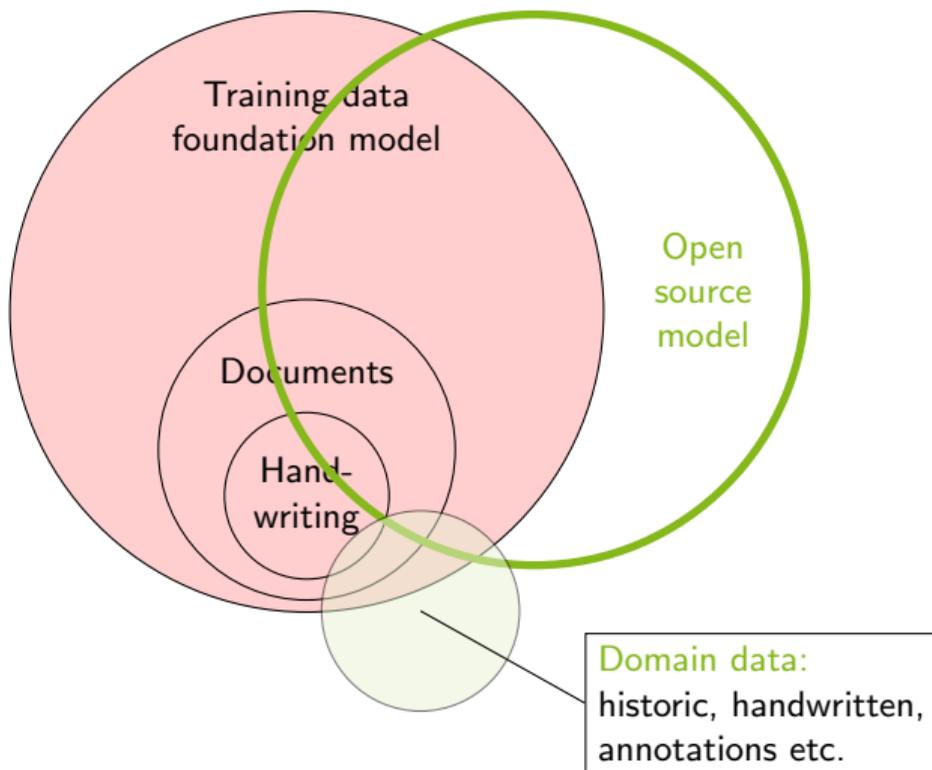
Problem #2: Insufficient Domain Coverage!



Problem #2: Insufficient Domain Coverage!



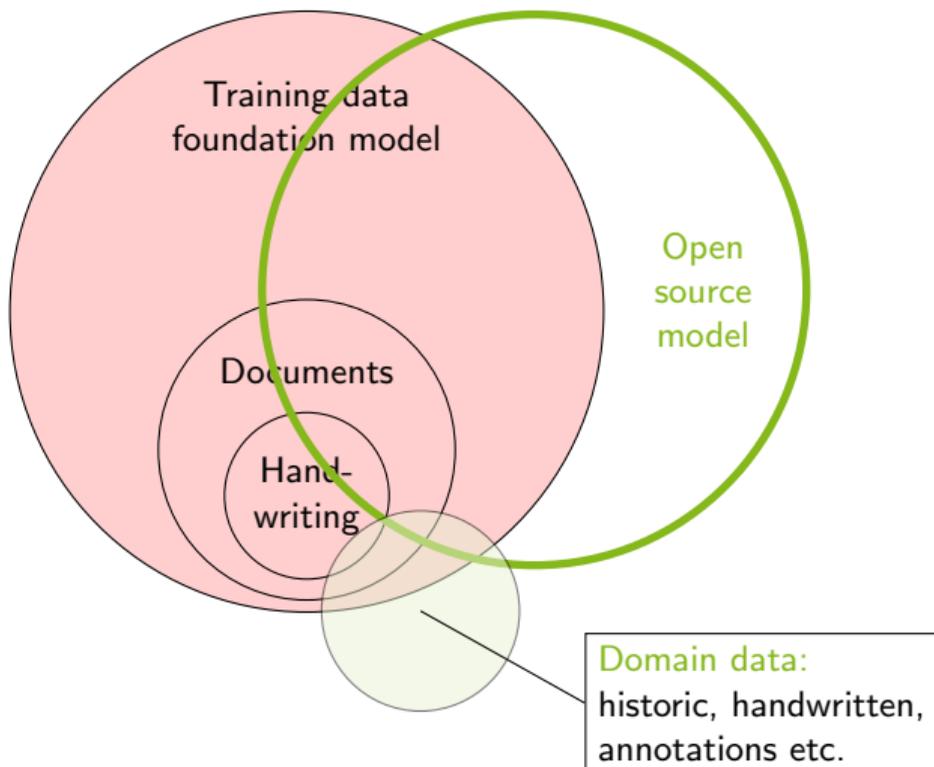
Problem #2: Insufficient Domain Coverage!



Problem #2: Insufficient Domain Coverage!

- ▶ Training data from the Internet
- ▶ Modern documents well covered
- ▶ Application domain hardly represented

- ▶ **Proposes solution:**
 - ▶ Adaptaion of the foundation models (data required!)
 - ▶ Expert models for specific tasks



Does that really work?

Pilot-Studies on CM/1 Files

Material: CM/1 Files (Germany)
(Cover sheets [“Mantelbögen” / 1st page of application])

Experiments in information extraction

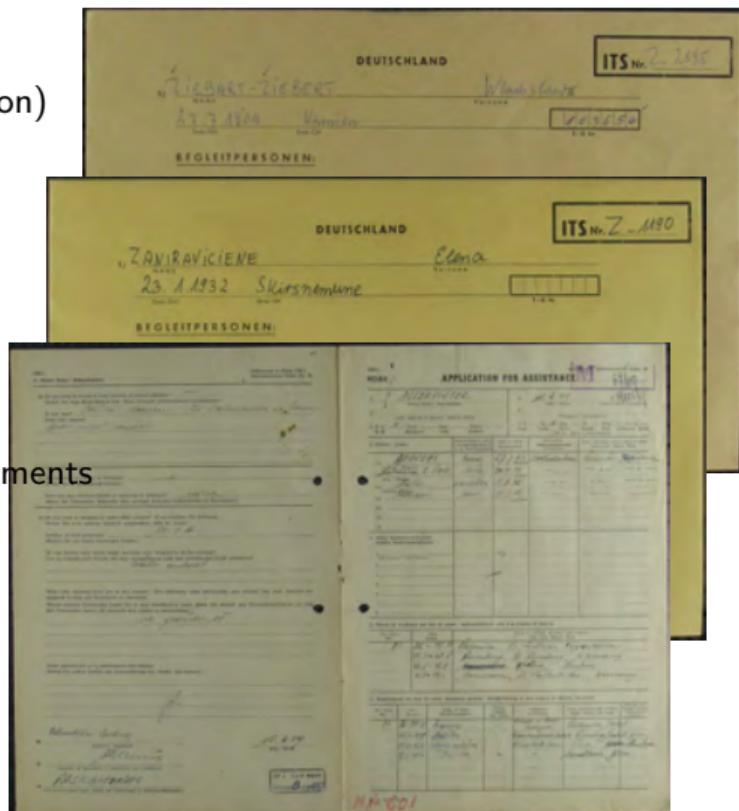
- ▶ All persons on the cover sheet (name, date of birth) in structured format (JSON)
- ▶ Information on main person of the application (Name, Nationality, Place-of-Birth ...)

Modells for *full-page information extraction*
esp. OpenSource Large Vision Language Models

- ▶ Training: up to 120,951 respectively 90,427 documents
- ▶ Test: 10,000 documents

Evaluation metrics:

- ▶ Structure: Tree-edit distance (TED)
- ▶ F1-Score: Information item extraction (F1)
- ▶ Transcription: Character-error rate (CER)



CM/1 Files: Cover Sheet

Dortmund

ITS Nr. L-9189

1) LUSTIGER Arno
Vorname

7.5.1924 205225
TID Nr.

Geb.-Dat. Geb.-Ort

BEGLEITPERSONEN:

2) LUSTIGER Gitla
Vorname

15.3.1896

Geb.-Dat. Geb.-Ort

3) LUSTIGER Helga
Vorname

23.11.1923

Geb.-Dat. Geb.-Ort

4) LUSTIGER Estera
Vorname

15.2.1931

Geb.-Dat. Geb.-Ort

CM/1	<u>1</u>			
DP/2				
Kr.-Pap.				

Krankens-Nr.:

{ "79429549": [

 { "Name": "LUSTIGER",

 "Vorname": "ARNO",

 "Geb-Dat": "1924-05-07" },

 { "Name": "LUSTIGER",

 "Vorname": "GITLA",

 "Geb-Dat": "1896-03-15" },

 { "Name": "LUSTIGER",

 "Vorname": "HELGA",

 "Geb-Dat": "1923-11-23" },

 { "Name": "LUSTIGER",

 "Vorname": "ESTERA",

 "Geb-Dat": "1931-02-15" }] } ;

¹CM/1 Files, Arolsen Archives <https://collections.arolsen-archives.org/de/archive/3-2-1-1>

CM/1 Files: Application Forms

APPLICATION FOR ASSISTANCE
FORO

NAME: [Lastname] [Firstname]

DATE: [Date]

Address: [Address]

Occupation: [Occupation]

Family: [Family]

Remarks: [Remarks]

Signature: [Signature]

Stamp: [Stamp]

Handwritten note: Mx-2325

APPLICATION FOR ASSISTANCE
FORO

NAME: [Lastname] [Firstname]

DATE: [Date]

Address: [Address]

Occupation: [Occupation]

Family: [Family]

Remarks: [Remarks]

Signature: [Signature]

Stamp: [Stamp]

AKERMAN

NAME: [Lastname] [Firstname]

DATE: [Date]

Address: [Address]

Occupation: [Occupation]

Family: [Family]

Remarks: [Remarks]

Signature: [Signature]

Stamp: [Stamp]

■ Firstname

■ Lastname

■ Date

¹CM/1 Files, Arolsen Archives <https://collections.arolsen-archives.org/de/archive/3-2-1-1>

Information Extraction on CM/1 Files: First Results

Persons

Model	1%		10%		25%		100%	
	TED	CER	TED	CER	TED	CER	TED	CER
Donut	25.6	53.5	88.2	8.9	90.4	6.9	92.4	6.9
PaliGemma	75.5	13.3	83.2	11.3	88.0	8.3	90.1	8.4
Qwen2.5-VL	58.8	15.9	87.0	9.1	89.3	8.1	89.8	7.8

Name

Model	1%		10%		25%		100%	
	Acc	CER	Acc	CER	Acc	CER	Acc	CER
Donut	0.0	93.5	0.0	85.1	53.5	15.1	61.1	12.3
PaliGemma	26.5	27.3	40.3	18.7	47.0	15.7	54.3	13.3
Qwen2.5-VL	35.1	20.8	46.0	17.0	53.8	13.8	55.2	12.5

F. Wolf, O. Tüselmann, A. Matei, L. Hennies, Ch. Rass, G. A. Fink: [CM1 — A Dataset for Evaluating Few-Shot Information Extraction with Large Vision Language Models](#), Proc. Int. Conf. on Document Analysis and Recognition (ICDAR), Wuhan, China, 2025, available here: arXiv.

CM/1 Files: Additional Fields

CM/1
PCIRO APPLICATION FOR ASSISTANCE

1. Name: Ackermann
2. Birth-date: 16.11.49
3. Birth-place: Wuerzburg
4. Nationality: Deutsch

5. Family Members of Family

No.	Name	Birth-date	Birth-place	Nationality
1	Fritz Ackermann	16.11.49	Wuerzburg	Deutsch
2	Lucie Ackermann	18.11.20	Lodz	Poln.
3	Viktor Ackermann	19.13.20	Guben	Deutsch.

6. Places of residence for last 12 years

Date	Place	Reason
1934-1939	Wuerzburg	Wuerzburg
1939-1941	Wuerzburg	Wuerzburg
1941-1942	Wuerzburg	Wuerzburg
1942-1943	Wuerzburg	Wuerzburg
1943-1944	Wuerzburg	Wuerzburg
1944-1945	Wuerzburg	Wuerzburg

7. Employment for last 12 years

Date	Place	Type of work	Employer	Reason for leaving
1934-1939	Wuerzburg	Wuerzburg	Wuerzburg	Wuerzburg
1939-1941	Wuerzburg	Wuerzburg	Wuerzburg	Wuerzburg
1941-1942	Wuerzburg	Wuerzburg	Wuerzburg	Wuerzburg
1942-1943	Wuerzburg	Wuerzburg	Wuerzburg	Wuerzburg
1943-1944	Wuerzburg	Wuerzburg	Wuerzburg	Wuerzburg
1944-1945	Wuerzburg	Wuerzburg	Wuerzburg	Wuerzburg

Handwritten notes: MX-2225, 5295

Legend

head-of-family

- name
- birth-date
- birth-place
- nationality

family-members

- name
- birth-date
- birth-place
- nationality

family-level

- religion
- family-name

```
{
  "head-of-family": {
    "name": "FRITZ ACKERMANN",
    "birth-date": "1881-01-05",
    "birth-place": "WUERZBURG",
    "nationality": "DEUTSCHER,
    HALBJUDE"
  },
  "family-members": [
    {
      "name": "LUCIE ACKERMANN",
      "birth-date": "1888-11-20",
      "birth-place": "LODZ",
      "nationality": "HALBJUDE,
      POLE"
    },
    {
      "name": "VIKTOR ACKERMANN",
      "birth-date": "1913-03-20",
      "birth-place": "GUBEN",
      "nationality": "DEUTSCHER,
      HALBJUDE"
    }
  ],
  "religion": "ROEMISCH-KATHOLISCH"
}
```

CM/1 Files: Additional Fields

CM/1
PCIRO
APPLICATION FOR ASSISTANCE

1. **Ackermann**
Family Name / Familienname

4. 16. 11. 49
Date / Datum

5. 967509
Identity No. / Ansichts-Nr.

2. *Other spelling or alias / Anderer Name*

6. Nationality / Nationalität
Deutsche

7. R. C. / D. K. / Prof. / Religion / Iron. / Job. / Other / Anders

7. S. / M. A. / Sep. / D. / W. L. / Un. C. / Leib. / Yrsh. / Ger. / Gsch. / W. / Barter South / Familienstand

8. Names / Name	Relationship to head / Verwandtschaftsverhältnis z. d. Familienhaupt	Date of birth / Geburtsdatum	Nationality / Staatsangehörigkeit / Nationalität	Town, province and country of birth / Geburtsort, Reich, Staat
H. Fsta - Karl	Haupt	5. 4. 81	Deutsche	Worpling / D.
Weyranh Annie	Wife	20. 11. 81	Poln. Kaiserin	Lady / Polen.
Weller Wenne	Child	20. 3. 13	Deutsche	Quibenz / D.

9. Other Members of Family / Andere Familienmitglieder

Legend

head-of-family

name

birth-date

birth-place

nationality

family members

¹Image source: CM/1 Files, Arolsen Archives <https://collections.arolsen-archives.org/de/archive/3-2-1-1>

Example: CM/1 Files

WUSA Form No. 1070-11 (1 Case 44)

NAME: LEIDA AALUND

ADDRESS: WUSA UNITED STATES ARMY

DATE: 10/10/50

REMARKS: LEIDA AALUND

SIGNATURE: [Signature]

name: LEIDA AALUND
nationality: ESTE
birth-place: TARTU

Annotation (top) / Prediction (bottom)

name: LEIDA AALUND
LEIDA AALUND
nationality: ESTE
ESTE
birth-place: TARTU
TARTU

¹Image source: CM/1 Files, Arolsen Archives <https://collections.arolsen-archives.org/de/archive/3-2-1-1>

Example: CM/1 Files

TUSA Form No. 2 629407 ESTONIAN
DP/O-1+ (1 June 46) HEADQUARTERS THIRD UNITED STATES ARMY
WARNA 190-265 KUSIMUSLEHT D.F.-dele (Umberpaiknenud isikutele)
DP Lager No. 615 Vanus 40 Sugu NAIS DP indeksekaardi No. 92778
Peajuhataoõrid: KÕIK KUSIMUSED PEAVAD OLEMA VASTATUD. VASTUSED OLGU KIRJUTATUD TRUKITAHTEDEGA. Kui teie vastused on mittetäielikud või valed, teid karistatakse USA sõjaväe seaduste alusel.
Kus ruum on varutud "jah" või "ei" jaoks, kasutage märki "X" oma vastuse näitamiseks.

1. Nimi (algupärase kirjaväe) AALUND LEIDA
(Perekonnanimi) (Eesnimi) (Nelupõlvnime)

2. a. Sünnikoht ja riik TARTU, ESTONIA b. Sünniea 21.10.1917
(päev, kuu, aasta)

3. a. Teie rahvus ESTLANC b. Teie usk LUTERI
c. Milline oli 31. augustil 1939 teie (seaduslik kodakondsus) EESTI
(aadress) TARTU, KASTANI NR 9-2

4. a. Teie nime täpne saksa keelne kirjaväe AALUND LEIDA
b. Nimetage kõik teised nimed, mille all olete või olete tuntud:

Nimi	Kus kasutatud	Kuupäev
EV OLE OLEKUD		

Annotation (top) / Prediction (bottom)

name: LEIDA AALUND

LEIDA AALUND

nationality: ESTE

ESTE

birth-place: TARTU

TARTU

¹Image source: CM/1 Files, Arolsen Archives <https://collections.arolsen-archives.org/de/archive/3-2-1-1>

Example: CM/1 Files

APPLICATION FOR ASSISTANCE

NAME: DELAMANIS
NATIONALITY: GREEK
BIRTH-PLACE: Athen, Greece

Year of Birth	Name	Relationship	Age	Sex	Marital Status	Occupation
18-41	Athen, Greece	41-41 - 41-41				
41-45	Maria, Greece					
45-47	Katerina, Greece					
47-48	Maria, Greece					
48-49	Katerina, Greece					

Annotation (top) / Prediction (bottom)

name: ISIDOR DELAMANIS

ISIDOR DELAMANIS

nationality: GRIECHE

GRIECHE

birth-place: ATHEN

ATHEN

¹Image source: CM/1 Files, Arolsen Archives <https://collections.arolsen-archives.org/de/archive/3-2-1-1>

Example: CM/1 Files

CM-1 APPLICATION FOR ASSISTANCE

1. Family Name / Familienname: DELAMANIS

2. Date / Datum: 19.7.49

3. Nationality / Nationalität: Greek

4. Name / Name: Isidor

5. Date of birth / Geburtsdatum: 24.4.30

6. Nationality / Nationalität: Greek

7. Place of birth / Geburtsort: Athen, Greece

Annotation (top) / Prediction (bottom)

name: ISIDOR DELAMANIS

ISIDOR DELAMANIS

nationality: GRIECHE

GRIECHE

birth-place: ATHEN

ATHEN

¹Image source: CM/1 Files, Arolsen Archives <https://collections.arolsen-archives.org/de/archive/3-2-1-1>

Example: CM/1 Files

Annotation (top) / Prediction (bottom)

name: JOSEF KISS
JOSEF KISS
nationality: UNGAR
UNGAR
birth-place: PELMONOSTOR
PELMONOSTOR

¹Image source: CM/1 Files, Arolsen Archives <https://collections.arolsen-archives.org/de/archive/3-2-1-1>

Example: CM/1 Files

E.C.No. 226 513

APPENDIX I to
Provisional Order No. 73

INTERNATIONAL REFUGEE ORGANIZATION
HEADQUARTERS GENEVA

Form No. CM/1 (Revised)

APPLICATION FOR IRO ASSISTANCE

(Bitte um IRO-Unterstützung)

1. *KISS*
Family Name in block capitals
(Familienname in großen Druckbuchstaben)

2. *832323*
(Other spellings or aliases by which family is known)
(Andere Schreibweisen oder Namen unter welchen die Familie bekannt ist)

3. Religion: (Check one with an "x")
 Roman Catholic (Römisch-katholisch)
 Protestant (evangelisch)
 Greek Orthodox (Griechisch-orthodox)
 Jewish (Jüdisch)
 Other (specify) (Andere (specifizieren))

4. (Country of last habitual residence prior to displacement)
(Letzter ständiger Wohnsitz vor Verdrängung)

5. (Date of completion of Form CM/1)
(Tag der Fertigstellung von Form CM/1)

6. Marital Status (Check one with an "x")
 Married (Verheiratet)
 Single (Ledig)
 Widowed (Witwenstand)
 Divorced (Geschieden)
 Unmarried (unmarried) (Unverheirateter Mann)

Name of all members of family living together	Relationship to head of family	Sex	Date of birth	Country of citizenship	Overseas ID No. (if any)	Town, province and country of birth
Namen aller Familienmitglieder wohnend im Zusammenleben	Verhältnis zum Familien-Vorstand	Geschlecht	Geburtsdatum	Staatsangehörigkeit	Autonummer im Ausl. ID (falls vorhanden)	Stadt, Landkreis und Staat der Geburt

Annotation (top) / Prediction (bottom)

name: JOSEF KISS
JOSEF KISS
nationality: UNGAR
UNGAR
birth-place: PELMONOSTOR
PELMONOSTOR

¹Image source: CM/1 Files, Arolsen Archives <https://collections.arolsen-archives.org/de/archive/3-2-1-1>

Example: CM/1 Files

Camp Estonia

RESETTLEMENT RECORD

Original Duplicate (Specify date for duplicate)

I. REGISTRANT									
Name (Surname)	Given name	Date of birth	Sex	Place of birth	Country of birth	Religion	Country of origin	Country of birth	Country of birth
GERDA	AASUMETS	1912	F	ESTONIA	ESTONIA				
II. DOCUMENTS (Under what name was he/she?)									
1. Passport (Date issued, valid to)									
2. Other documents (Date issued, valid to)									
III. PREVIOUS RESIDENCE					IV. DATE TO WHICH TO PREVIOUS RESIDENCE				
Country					Date				
Country: <i>Estonia</i>									
V. EDUCATION									
Country									
Country: <i>Estonia</i>									
VI. TRADE AND EMPLOYMENT CLASSIFICATION									
Industry									
Occupation									
Occupation: <i>Dec. 1-2</i>									
VII. THREE COUNTRIES OF PRECEDENCE FOR RESETTLEMENT									
Country									
Country: <i>Estonia, Latvia, Lithuania</i>									
VIII. PREVIOUS APPLICATION TO EMIGRATE									
Authority to which application made									
Date of last application									
Country to which visa or permission is held, if held									
IX. LOCATIONS									
Date									
Place									
Place: <i>Estonsk, Latvia, Lithuania</i>									

U.S. Census Bureau

Annotation (top) / Prediction (bottom)

name: GERDA AASUMETS

GERDA AASUHETS

nationality: null

null

birth-place: PALMSE

PALMSE

¹Image source: CM/1 Files, Arolsen Archives <https://collections.arolsen-archives.org/de/archive/3-2-1-1>

Example: CM/1 Files

Camp Estonia

RESETTLEMENT RECORD

Original Duplicate IDENTITY CARD No. *214982*

1. REGISTRANT										
Family Surname <i>AASUMETS</i>	Other names <i>GERDA</i>		Place of Birth	Sex	Married	Widowed	Single	Divorced	Other	Designation
Date of Birth <i>19.2.1922</i>	Place of Birth <i>PALMSE/ESTONIA</i>	Province <i>ESTONIA</i>	Country <i>ESTONIA</i>	Religion <i>ESTONIAN</i>	Citizenship <i>ESTONIAN</i>	Country of Birth				
2. DOCUMENTS (Strike out those not held: PASSPORT, BIRTH CERTIFICATE, MARRIAGE CERTIFICATE)										
3. PREVIOUS RESIDENCE			4. DESIRE TO RETURN TO PREVIOUS RESIDENCE							
5. EDUCATION										
6. TRADE AND EMPLOYMENT CLASSIFICATION										
7. THREE COUNTRIES OF PREFERENCE FOR RESETTLEMENT										

Annotation (top) / Prediction (bottom)

name: GERDA AASUMETS

GERDA AASUHETS

nationality: null

null

birth-place: PALMSE

PALMSE

¹Image source: CM/1 Files, Arolsen Archives <https://collections.arolsen-archives.org/de/archive/3-2-1-1>

Example: CM/1 Files

Annotation (top) / Prediction (bottom)

name: HRYHORIJ DUDTSCHENKO
nationality: POLE, UKRAINER
birth-place: PIDBORCI

¹Image source: CM/1 Files, Arolsen Archives <https://collections.arolsen-archives.org/de/archive/3-2-1-1>

Example: CM/1 Files

Form No. 100-10
100-10
100-10

HEADQUARTERS
THIRD UNITED STATES ARMY

BERNHARD DERKSEN

Do Not Write Here

NAME: BERNHARD DERKSEN

NATIONALITY: MENNONIT

BIRTH-PLACE: OSTERWICK MOSKAU

Annotation (top) / Prediction (bottom)

name: BERNHARD DERKSEN
KATHARINA BERKSON

nationality: MENNONIT
NANSEN-PASS, RUSSE, STAATENLOS

birth-place: OSTERWICK
MOSKAU

¹Image source: CM/1 Files, Arolsen Archives <https://collections.arolsen-archives.org/de/archive/3-2-1-1>

Example: CM/1 Files

Annotation (top) / Prediction (bottom)

name: JOZEF PELECHATY
JOZEF PELECHATA

nationality: POLE
POLE

birth-place: DZWINAGROD BUCZACZ
DZWINOGROD

¹Image source: CM/1 Files, Arolsen Archives <https://collections.arolsen-archives.org/de/archive/3-2-1-1>

Information Extraction on CM/1 Files: Latest Results

Single and Multi-Field Extraction

Model	Fields	1%		10%		25%		100%	
		CER ↓	F1 ↑						
PaliGemma	Name	33.71	18.64	21.63	29.83	17.09	38.74	10.64	56.69
	Nationality (Nat)	36.90	57.03	23.98	75.38	19.91	78.54	16.62	82.31
	Place of Birth (PoB)	59.28	15.39	45.59	28.87	40.23	30.62	33.18	40.05
	Religion	47.54	58.84	27.31	78.50	20.39	82.82	18.85	84.47
	Name,PoB	38.03	22.75	27.17	34.78	22.90	43.45	17.12	55.81
	Name,Nat,PoB	38.62	42.68	25.82	56.19	21.49	62.84	16.46	70.63
Donut	Name	51.22	8.01	26.61	31.56	17.79	44.79	9.90	61.96
	Nationality (Nat)	40.73	57.64	23.79	76.24	20.63	79.54	15.60	84.06
	Place of Birth (PoB)	56.56	18.49	41.20	31.83	36.18	38.20	29.86	46.07
	Religion	67.87	44.80	29.53	78.62	20.44	83.93	16.48	86.14
	Name,PoB	51.22	16.17	26.85	40.76	21.57	51.09	15.53	61.67
	Name,Nat,PoB	50.47	35.39	25.17	60.41	20.46	66.66	14.83	74.43

A. Matei, T. Hallyburton, L. Hennies, Ch. Rass, G. A. Fink: [Recent Advances in Information Extraction from Historical Archival Records](#), Proc. Int. Conf. on Document Analysis and Recognition (ICDAR), Vienna, Austria, 2026, *under review*.

Information Extraction on CM/1 Files: Latest Results

Isolated vs. Combined Extraction (Nationality)

Model	Fields	1%	10%	25%	100%
PG	Nat (baseline)	57.03	75.38	78.54	82.31
	Name,Nat,PoB	60.61	74.63	80.57	84.20
D.	Nat (baseline)	57.64	76.24	79.54	84.06
	Name,Nat,PoB	51.76	75.61	79.87	84.11

Extraction (Nationality) After Pre-Training on Related Fields

Model	Fields	1%	10%	25%
PG	Nat (baseline)	57.03	75.38	78.54
	Name,PoB	68.80	78.53	81.00
D.	Nat (baseline)	57.64	76.24	79.54
	Name,PoB	70.75	79.03	80.73

A. Matei, T. Hallyburton, L. Hennies, Ch. Rass, G. A. Fink: [Recent Advances in Information Extraction from Historical Archival Records](#), Proc. Int. Conf. on Document Analysis and Recognition (ICDAR), Vienna, Austria, 2026, *under review*.

Summary

- ▶ *Size matters!*
Or: Automation of deep indexing of historic mass files is *without alternative!*
- ▶ Large Vision Language Models are a *powerful* future technology
- ▶ Models need to be adapted to the special problem (historic documents & handwriting)
- ⊘ How to avoid annotation effort for adaptation data?
- ⊘ Quality control is an unsolved problem!
- ⊘ Computational requirements *tremendous!*



Image Source: Census 1950, LAV NRW; Gernot A. Fink