European Centre for the Development of Vocational Training

50 YEARS
SHAPING LEARNING AND SKILLS FOR EUROPE

Using language models for extracting regions of employment from online job vacancies

Adam Tsakalidis & Antonio Ranieri

Web Intelligence Network Conference – From Web to Data

Adam Tsakalidis
Cedefop Expert (Skills Intelligence & Foresight)

Gdańsk, Poland I 04 Feb 2025

Introduction

- **Task**: Extracting regions of employment from Online Job Advertisements (OJAs)

Introduction

- **Task**: Extracting regions of employment from Online Job Advertisements (OJAs)

- **Why this matters**
  - OJAs: ‘real-time’ monitoring of the labour market
  - Regions: fine-grained resolution

Introduction

- **Task**: Extracting regions of employment from Online Job Advertisements (OJAs)

- **Why this matters**
  - OJAs: ‘real-time’ monitoring of the labour market
  - Regions: fine-grained resolution

- **Data:** Skills-OVATE
  - ESTAT + Cedefop
  - Hundreds of millions of OJAs
  - Multiple languages
  - EU27+ coverage
  - 2018-24
  - Classifications: occupations (ISCO), skills (ESCO), …

Introduction

- **Task**: Extracting regions of employment from Online Job Advertisements (OJAs)

- **Why this matters**
  - OJAs: ‘real-time’ monitoring of the labour market
  - Regions: fine-grained resolution

- **Data:** Skills-OVATE
  - ESTAT + Cedefop
  - Hundreds of millions of OJAs
  - Multiple languages
  - EU27+ coverage
  - 2018-24
  - Classifications: occupations (ISCO), skills (ESCO), …

In this work:
- Greek language
- NUTS-2 region

Challenges

- **Plan A:** Using commercial LLMs

- Data sensitivity
- Cost

Challenges

- **Plan A**: Using commercial LLMs
- **Plan B**: Prompting in-house LLMs

- Cost
- Scalability

Challenges

- **Plan A**: Using commercial LLMs
- **Plan B**: Prompting in-house LLMs
- **Plan C**: Typical ML/NLP process
  - (a) Manual annotation of two datasets (train/test)
  - (b) Training a model on the training data
  - (c) Evaluate model on the test data

Challenges

- **Plan A**: Using commercial LLMs
- **Plan B**: Prompting in-house LLMs
- **Plan C**: Typical ML/NLP process
  - (a) Manual annotation of two datasets (train/test)
  - (b) Training a model on the training data
  - (c) Evaluate model on the test data

- Time consuming (a)
- Labour intensive (a)

Challenges

- **Plan A**: Using commercial LLMs
- **Plan B**: Prompting in-house LLMs
- **Plan C**: Typical ML/NLP process
  - (a) Manual annotation of two datasets (train/test)
  - (b) Training a model on the training data
  - (c) Evaluate model on the test data
- **Summary of challenges**
  - Privacy-preserving
  - Low cost
  - Scalable

Challenges

- **Plan A**: Using commercial LLMs
- **Plan B**: Prompting in-house LLMs
- **Plan C**: Typical ML/NLP process
  - (a) Manual annotation of two datasets (train/test)
  - (b) Training a model on the training data
  - (c) Evaluate model on the test data
- **Summary of challenges**
  - Privacy-preserving
  - Low cost
  - Scalable

Overview of our approach

Overview of our approach

Step 1: Location candidates

OJA-1: SSP ltd, a thriving force in online marketing, with offices in **Athens**, **Mykonos** and **Crete**, is looking for an enthusiastic mid-level Java developer for our new office in **Thessaloniki**. If interested, send your CV to yptoleme@myemail.gr.

OJA-2: We are looking for a chef for our brand new Hotel in **Santorini** to join our team for the summer of 2024. Main responsibilities: <...> Requirements: <...>

OJA-3: Our client, a growing startup based in **Patra**, is looking for a new HR manager to join our team [...] Main responsibilities: <...> Requirements: <...>

Overview of our approach

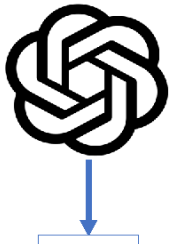
Step 1: Location candidates

Step 2: LLMs as ‘Experts’

OJA-1: SSP ltd, a thriving force in online marketing, with offices in **Athens**, **Mykonos** and **Crete**, is looking for an enthusiastic mid-level Java developer for our new office in **Thessaloniki**. If interested, send your CV to yptoleme@myemail.gr.

What is the NUTS-2 region of the following word/phrase?

Athens



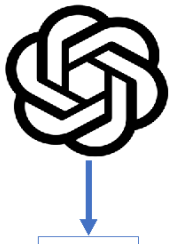
EL30

OJA-2: We are looking for a chef for our brand new Hotel in **Santorini** to join our team for the summer of 2024. Main responsibilities: <...> Requirements: <...>

• • •

What is the NUTS-2 region of the following word/phrase?

Patra



EL63

OJA-3: Our client, a growing startup based in **Patra**, is looking for a new HR manager to join our team [...] Main responsibilities: <...> Requirements: <...>

Overview of our approach

Step 1: Location candidates

Step 2: LLMs as 'Experts'

Step 3: Assigning NUTS-2 regions

OJA-1: SSP ltd, a thriving force in online marketing, with offices in **Athens**, **Mykonos** and **Crete**, is looking for an enthusiastic mid-level Java developer for our new office in **Thessaloniki**. If interested, send your CV to yptoleme@myemail.gr.

What is the NUTS-2 region of the following word/phrase?

Athens

EL30

OJA-1: SSP ltd, a thriving force in online marketing, with offices in **Athens**, **Mykonos** and **Crete**, is looking for an enthusiastic mid-level Java developer for our new office in **Thessaloniki**. If interested, send your CV to yptoleme@myemail.gr.

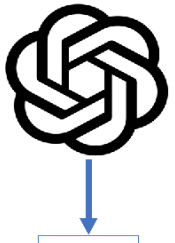
Labels: EL30 EL42 EL43 EL52

OJA-2: We are looking for a chef for our brand new Hotel in **Santorini** to join our team for the summer of 2024. Main responsibilities: <...> Requirements: <...>

• • •

What is the NUTS-2 region of the following word/phrase?

Patra



EL63

OJA-2: We are looking for a chef for our brand new Hotel in **Santorini** to join our team for the summer of 2024. Main responsibilities: <...> Requirements: <...>

Labels: EL42

OJA-3: Our client, a growing startup based in **Patra**, is looking for a new HR manager to join our team [...] Main responsibilities: <...> Requirements: <...>

OJA-3: Our client, a growing startup based in **Patra**, is looking for a new HR manager to join our team [...] Main responsibilities: <...> Requirements: <...>

Labels: EL63

Step 1: Extract candidate location terms

- **Pre-defined lists**
  - Non-contextual → false positives

Step 1: Extract candidate location terms

- **Pre-defined lists**
  - Non-contextual → false positives

- **Named entity recognition**
  - Language-dependent

Step 1: Extract candidate location terms

- **Pre-defined lists**
  - Non-contextual → false positives
- **Named entity recognition**
  - Language-dependent
- **Embeddings-based approach**

Step 1: Extract candidate location terms

- **Pre-defined lists**
  - Non-contextual → false positives
- **Named entity recognition**
  - Language-dependent
- **Embeddings-based approach**
  - Anchors: manually created list of location names (est. effort: 5')

Step 1: Extract candidate location terms

- **Pre-defined lists**
  - Non-contextual → false positives
- **Named entity recognition**
  - Language-dependent
- **Embeddings-based approach**
  - Anchors: manually created list of location names (est. effort: 5')
  - Anchors contextual embeddings: extracted from OJAs (Greek language)

Step 1: Extract candidate location terms

- **Pre-defined lists**
  - Non-contextual → false positives
- **Named entity recognition**
  - Language-dependent
- **Embeddings-based approach**
  - Anchors: manually created list of location names (est. effort: 5')
  - Anchors contextual embeddings: extracted from OJAs (Greek language)
  - **Candidates**: highly similar embeddings to anchors

Overview of our approach

Step 1: Location candidates

Step 2: LLMs as 'Experts'

Step 3: Assigning NUTS-2 regions

OJA-1: SSP ltd, a thriving force in online marketing, with offices in **Athens**, **Mykonos** and **Crete**, is looking for an enthusiastic mid-level Java developer for our new office in **Thessaloniki**. If interested, send your CV to yptoleme@myemail.gr.

What is the NUTS-2 region of the following word/phrase?

Athens

EL30

OJA-1: SSP ltd, a thriving force in online marketing, with offices in **Athens**, **Mykonos** and **Crete**, is looking for an enthusiastic mid-level Java developer for our new office in **Thessaloniki**. If interested, send your CV to yptoleme@myemail.gr.

Labels: EL30 EL42 EL43 EL52

OJA-2: We are looking for a chef for our brand new Hotel in **Santorini** to join our team for the summer of 2024. Main responsibilities: <...> Requirements: <...>

• • •

What is the NUTS-2 region of the following word/phrase?

Patra

EL63

OJA-2: We are looking for a chef for our brand new Hotel in **Santorini** to join our team for the summer of 2024. Main responsibilities: <...> Requirements: <...>

Labels: EL42

OJA-3: Our client, a growing startup based in **Patra**, is looking for a new HR manager to join our team [...] Main responsibilities: <...> Requirements: <...>

OJA-3: Our client, a growing startup based in **Patra**, is looking for a new HR manager to join our team [...] Main responsibilities: <...> Requirements: <...>

Labels: EL63

Step 2: LLMs as ‘experts’

You are an NLP annotator. You are being provided with a phrase (unigram or bigram) in Greek, which is very likely to be the name of a location (not necessarily in Greece).

Your task is to find and output the NUTS-2 region of the phrase (or of part of the phrase) in the following format: <NUTS2_CODE><NUTS2_NAME> (e.g.: <EL30><Περιφέρεια Αττικής>)

If the phrase does not match any location, output: <NONE><NONE>

If the phrase is more generic (e.g., 'Greece'), output the respective NUTS-1 region, if available (or country code, otherwise): <NUTS1_CODE><NUTS1_NAME>

If you find multiple locations for the same phrase, list up to 3 locations:

<NUTS2_CODE1><NUTS2_NAME1>, …, <NUTS2_CODE3><NUTS2_NAME3>

Try to associate a phrase with a NUTS-2 code, even if you are not certain about it (avoid <NONE><NONE> as much as possible). Do not reason about your answers. The phrase is:

Overview of our approach

Step 1: Location candidates

Step 2: LLMs as 'Experts'

Step 3: Assigning NUTS-2 regions

OJA-1: SSP ltd, a thriving force in online marketing, with offices in **Athens**, **Mykonos** and **Crete**, is looking for an enthusiastic mid-level Java developer for our new office in **Thessaloniki**. If interested, send your CV to yptoleme@myemail.gr.

What is the NUTS-2 region of the following word/phrase?

Athens

EL30

OJA-1: SSP ltd, a thriving force in online marketing, with offices in **Athens**, **Mykonos** and **Crete**, is looking for an enthusiastic mid-level Java developer for our new office in **Thessaloniki**. If interested, send your CV to yptoleme@myemail.gr.

Labels: EL30 EL42 EL43 EL52

OJA-2: We are looking for a chef for our brand new Hotel in **Santorini** to join our team for the summer of 2024. Main responsibilities: <...> Requirements: <...>

• • •

What is the NUTS-2 region of the following word/phrase?

Patra

EL63

OJA-2: We are looking for a chef for our brand new Hotel in **Santorini** to join our team for the summer of 2024. Main responsibilities: <...> Requirements: <...>

Labels: EL42

OJA-3: Our client, a growing startup based in **Patra**, is looking for a new HR manager to join our team [...] Main responsibilities: <...> Requirements: <...>

OJA-3: Our client, a growing startup based in **Patra**, is looking for a new HR manager to join our team [...] Main responsibilities: <...> Requirements: <...>

Labels: EL63

LM fine-tuning

- **Step 3: Training dataset creation**
  - Map candidates/NUTS-2 labels from Step 2 back to their OJA descriptions
  - Keep OJAs with only one NUTS-2 region assigned
  - Result: **6,501** (OJA_text, NUTS2_region) tuples (from 10k OJAs)

LM fine-tuning

- **Step 3: Training dataset creation**
  - Map candidates/NUTS-2 labels from Step 2 back to their OJA descriptions
  - Keep OJAs with only one NUTS-2 region assigned
  - Result: **6,501** (OJA_text, NUTS2_region) tuples (from 10k OJAs)

- **Test dataset creation**
  - Manual annotation: **528** (OJA_text, NUTS2_region) written in Greek

LM fine-tuning

- **Step 3: Training dataset creation**
  - Map candidates/NUTS-2 labels from Step 2 back to their OJA descriptions
  - Keep OJAs with only one NUTS-2 region assigned
  - Result: **6,501** (OJA_text, NUTS2_region) tuples (from 10k OJAs)

- **Test dataset creation**
  - Manual annotation: **528** (OJA_text, NUTS2_region) written in Greek

- **Fine-tune BERT** (Koutsikakis et al., 2020) for classification task**:**
  - Train on the 6,501 OJA descriptions
  - Apply model on the 528 OJA descriptions
  - Measure accuracy on the 528 OJAs

Image source: github

Koutsikakis, John, et al. "Greek-bert: The greeks visiting sesame street." *11th Hellenic conference on artificial intelligence*. 2020.

Results

	Model Explanation
Majority	Predicts ‘EL30’ all the time
ChatGPT	Labels ‘candidate’ location terms extracted from each OJA in the test set. The output is the final prediction for each OJA.
Ours	A BERT-based model that is trained on 6,501 OJAs that were previously annotated by ChatGPT.

Results

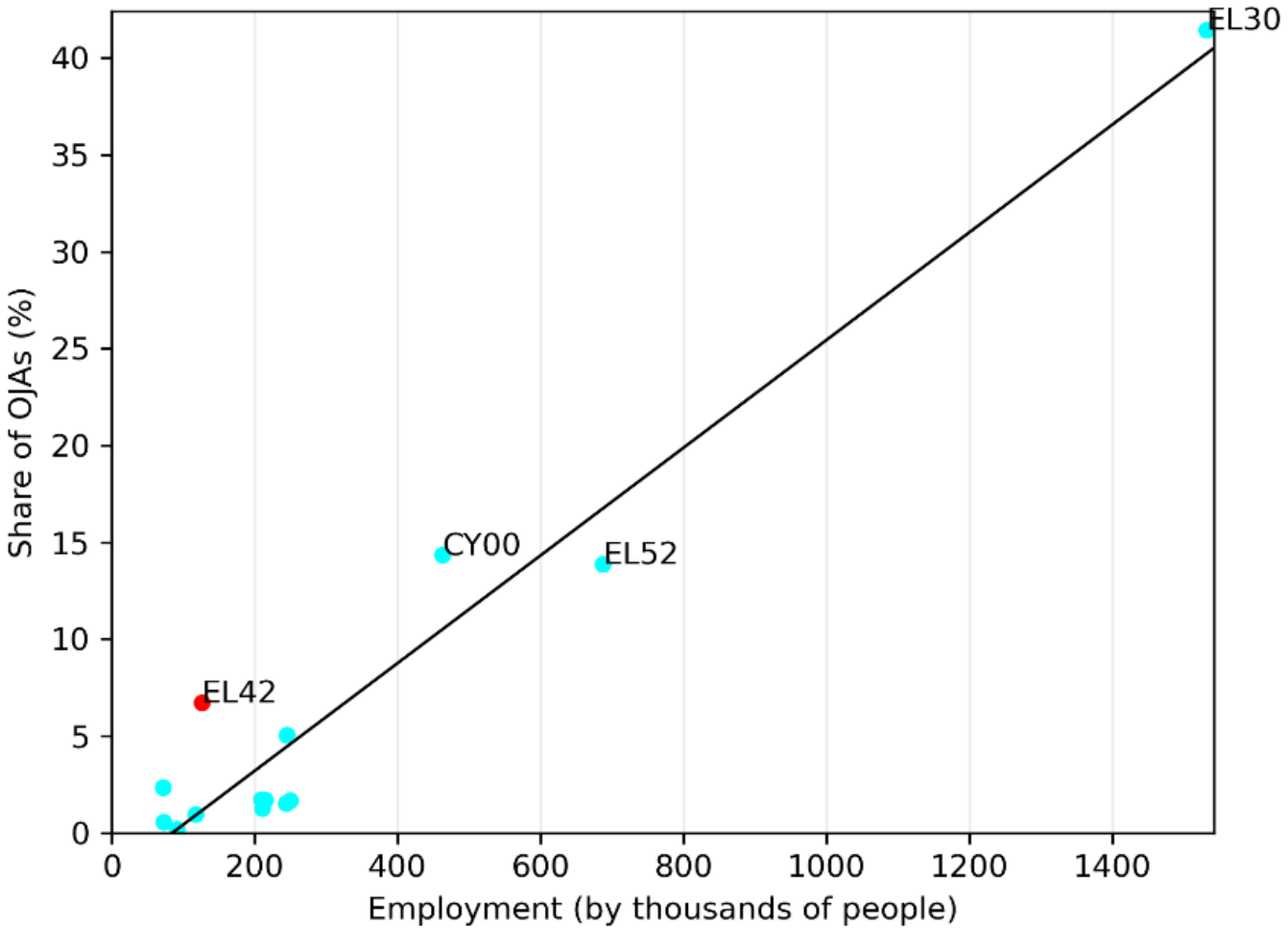
	Model Explanation	Incl. N/A regions	Excl. N/A regions
Majority	Predicts ‘EL30’ all the time	.254	.329
ChatGPT	Labels ‘candidate’ location terms extracted from each OJA in the test set. The output is the final prediction for each OJA.	**.813**	.828
Ours	A BERT-based model that is trained on 6,501 OJAs that were previously annotated by ChatGPT.	.767	**.933**

Results

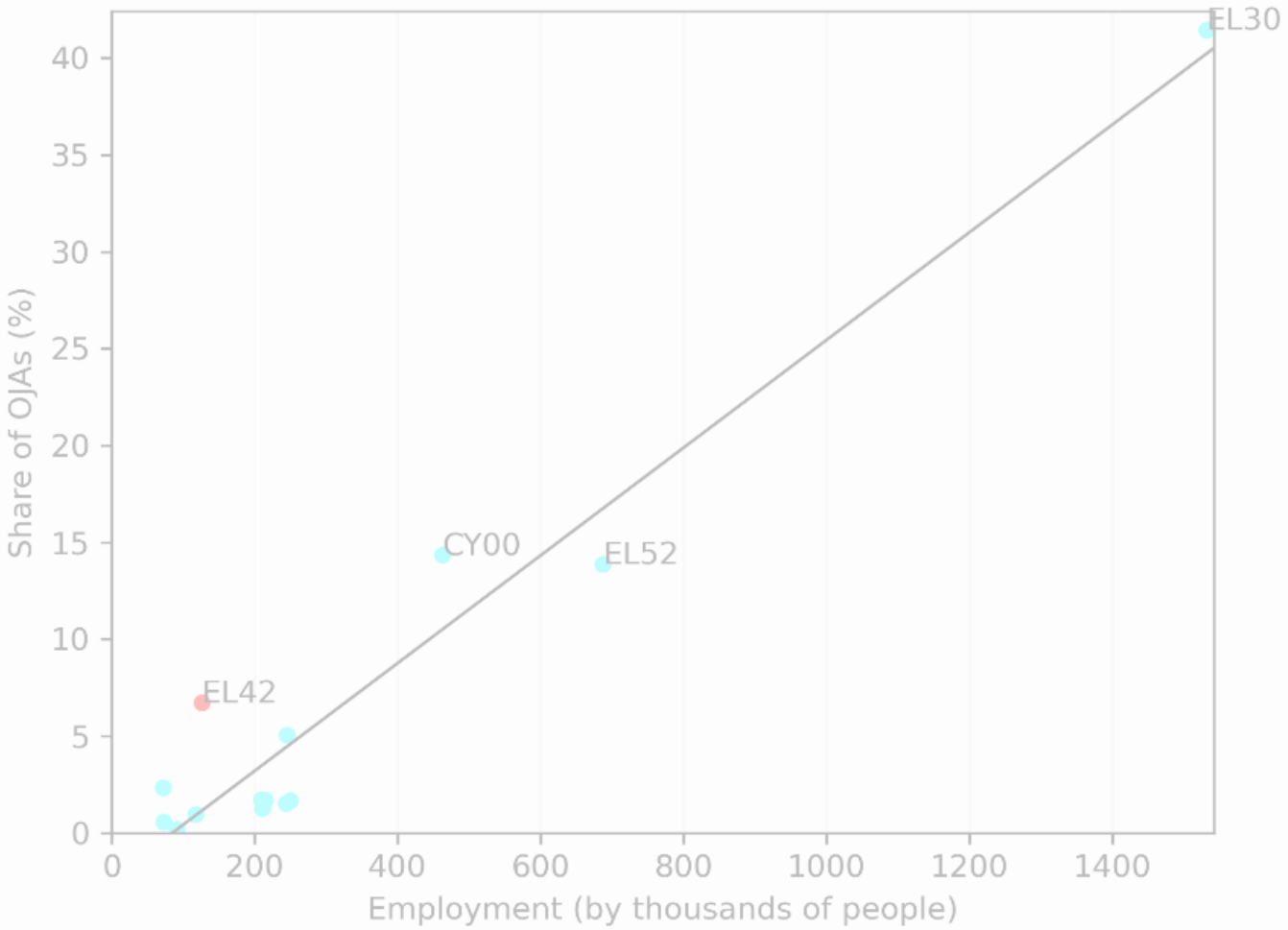
	Model Explanation	Incl. N/A regions	Excl. N/A regions	Cost
Majority	Predicts ‘EL30’ all the time	.254	.329	-
ChatGPT	Labels ‘candidate’ location terms extracted from each OJA in the test set. The output is the final prediction for each OJA.	.813	.828	**$0.25 (for 528 OJAs)**
Ours	A BERT-based model that is trained on 6,501 OJAs that were previously annotated by ChatGPT.	.767	.933	**$34 (for all OJAs in Greek)**

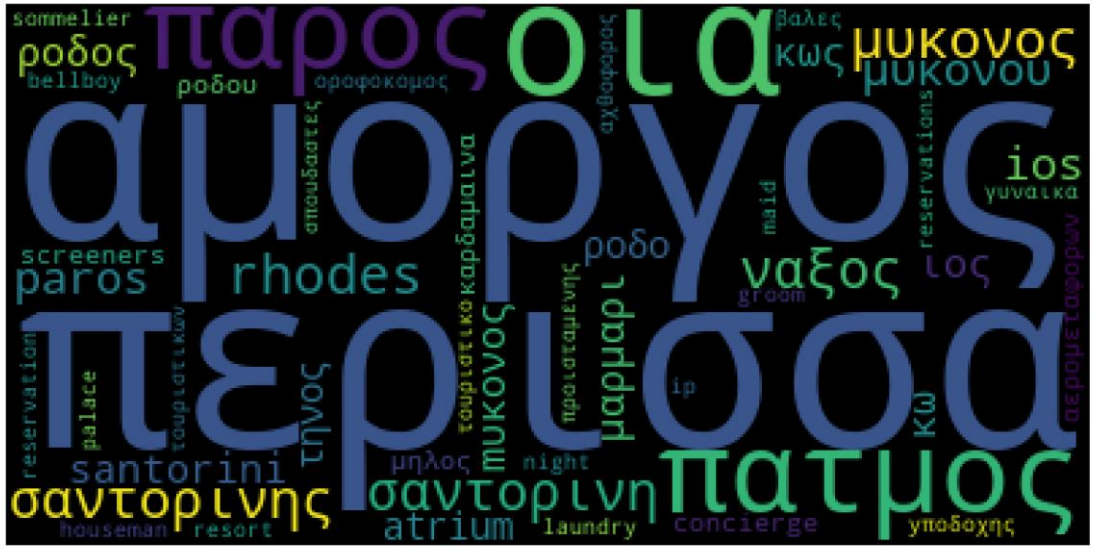
Correlation against official statistics & discussion

Correlation against official statistics & discussion

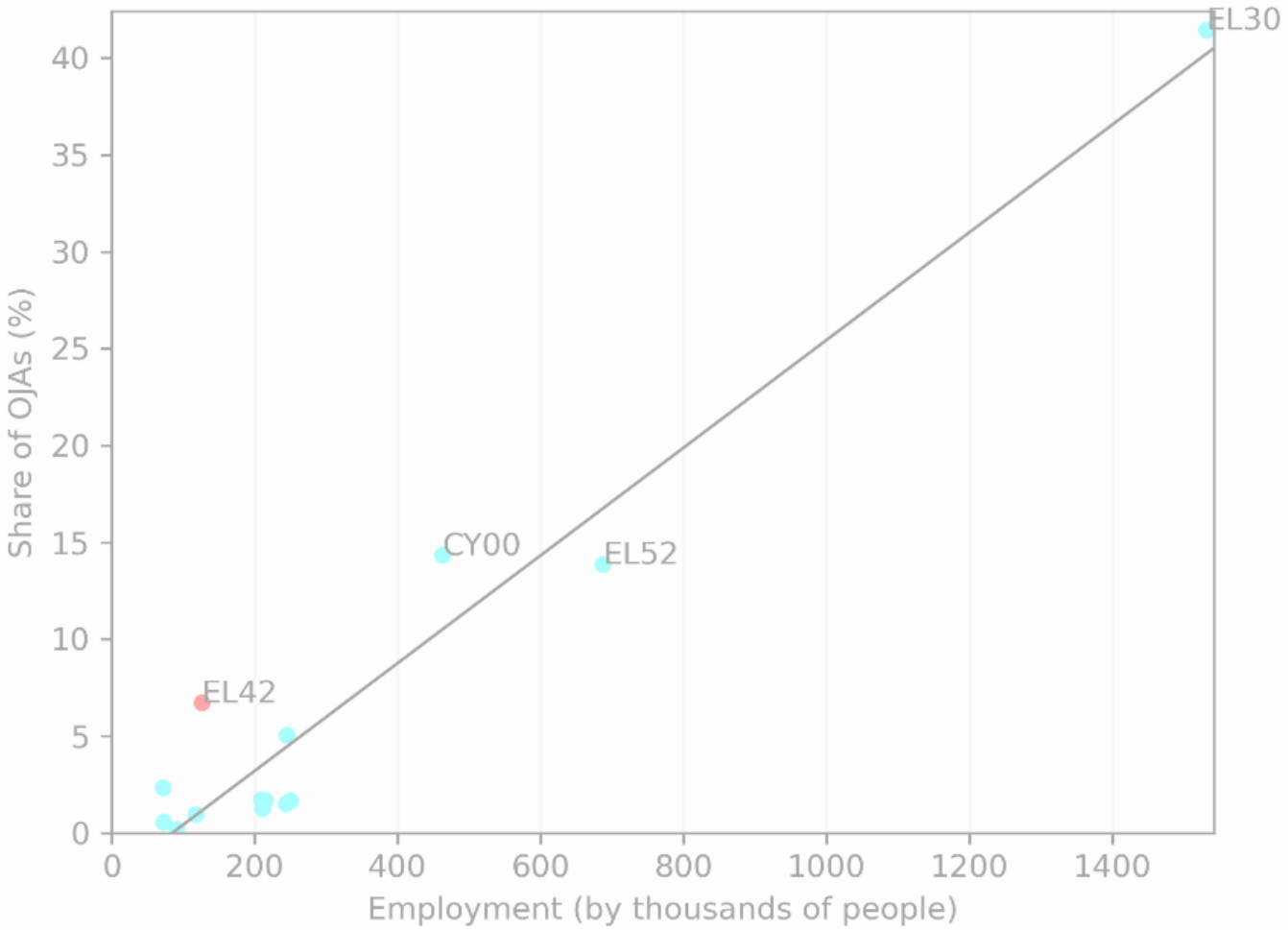


Correlation against official statistics & discussion





Correlation against official statistics & discussion



Most characteristic keywords:

palace, groom, laundry, reservations, porter, valet, touristic, sommelier, supervisor, reservation, houseman, students, resort, ip, bellboy, woman, night, reception, maid, airlines, screeners, concierge, atrium

Conclusion

- Task: extract NUTS-2 regions of employment from OJAs in Greek
  - LLMs as ‘annotators’
  - Extremely limited manual effort
  - Very high accuracy
  - Privacy-preserving approach
  - Scalability

Conclusion

- Task: extract NUTS-2 regions of employment from OJAs in Greek
  - LLMs as ‘annotators’
  - Extremely limited manual effort
  - Very high accuracy
  - Privacy-preserving approach
  - Scalability

- Limitations
  - Single language
  - Small test set

Conclusion

- Task: extract NUTS-2 regions of employment from OJAs in Greek
  - LLMs as ‘annotators’
  - Extremely limited manual effort
  - Very high accuracy
  - Privacy-preserving approach
  - Scalability

- Limitations
  - Single language
  - Small test set

- Future work
  - Expand the approach across languages
  - Robust validation

Thank you

www.cedefop.europa.eu

Follow us on social media

@adtsakal

Additional Links:

Skills OVATE
Cedefop Skills Intelligence

50 YEARS
SHAPING LEARNING AND SKILLS FOR EUROPE

CEDEFOP
European Centre for the Development of Vocational Training