A Hybrid Semantic Normalization Framework

for Mapping Multilingual Job Postings to European Taxonomies

P. Zervas, G. Tzimas, K. Giotopoulos (Univ. of Peloponnese)
Ł. Sienkiewicz (Gdańsk University of Technology) · F. Pesce (IRS, Bologna)

Cedefop × Eurostat — Harnessing Web Data for Next-Generation Skills Intelligence
Thessaloniki · 28–29 May 2026
EU-ALMPO project · Grant Agreement No. 101178736

Funded by
the European Union

Presentation Outline

- **EU-ALMPO project overview:** building an AI-supported Observatory for labour-market policy
- **Main challenge:** online job advertisements are heterogeneous, noisy, and difficult to compare across platforms, countries, languages, and sources.
- **The goal:** transform raw job-ad text into standardized, comparable labour-market data using European occupational and skills taxonomies.
- **The approach:** combine semantic search, rules, and controlled AI support
- **Expected outputs:** structured data for analytics, dashboards, and policy oriented tools
- **Evidence and Takeaways:** evaluation examples, lessons learned, and takeaways

The EU-ALMPO project

EU Active Labour Market Policies Observatory

A **Horizon Europe** project building an AI-supported observatory that translates fragmented labour market data into **actionable intelligence for ALMP design and evaluation** across EU member states.

Establish a European Observatory for Active Labour Market Policies — centralised, cross-national, evidence-based.

Build an AI-powered platform: knowledge base, design wizard, and labour market analytics for policymakers, researchers and PES professionals.

Target labour market integration challenges — with a focus on vulnerable groups and cross-country comparability.

EU CONTRIBUTION
€2.48M
Horizon Europe · GA 101178736

DURATION
Jan 2025 – Dec 2027
3-year project

CONSORTIUM
8 partners · 7 countries
Coordinator: University of Peloponnese

PILOT COUNTRIES
DK · GR · IT · ES
Administrative data + OJA feeds

The EU-ALMPO consortium

Coordinator: University of Peloponnese (GR) · 8 participants + 4 partners · 7 countries

University of Peloponnese
GREECE — COORDINATOR

Project lead, AI & NLP, normalization engineering, Data Science, Information Systems

Gdansk Univ. of Technology
POLAND

Labour-market research, PES

Inst. for Social Research
ITALY — IRS BOLOGNA

Labour-market research, policy evaluation

Tavistock Institute
GERMANY

Labour-market research, evaluation

DomSpain
SPAIN

Training & VET, dissemination

Acronym
Greece

Labour-market research, PES

HELLENIC ADULT EDUCATION ASSOCIATION (HAEA) |
Greece

Public employment service

p-consulting
GREECE

IT delivery, systems integration

ASSOCIATED PARTNERS

3F (DK) · INAPP (IT) · SDE Korydallos (GR) · SDE Peiraia (GR)

ADVISORY BOARD

Eurofound · OECD · World Bank — pressure-testing technical and policy choices.

EU-ALMPO · the observatory at a glance

01 SOURCES

S1 · CURATED LITERATURE

ALMP knowledge sources

Research papers, evaluation reports, policy manuals and ALMP design documents.

S2 · DEMAND SIGNALS

Online job postings

47 portals · 10+ countries · ~2–3M postings crawled for skills, occupations, sectors and trends.

S3 · ADMINISTRATIVE DATA

National PES & statistics agencies

Employed & unemployed populations, labour force participation and ALMP programme records from national PES systems & statistical offices.

02 ALMP HUB · TOOLS, DATA & BI

ALMP Hub · *role-based portal & AI-powered BI*

One portal for all user roles

Tool 1 · KNOWLEDGE

Curated ALMP Knowledge Base

Searchable repository enriched with metadata, evidence tags and AI Q&A (RAG). Annotator tool with human expert review.

Tool 2 · DESIGN

ALMP Design Wizard

Agentic AI that guides stakeholders step-by-step in designing ALMPs, drawing on all sources. Labour-market signals & tailored recommendations.

T3 · LM OBSERVATORY DASHBOARD

Labour-market analytics & dashboards

Skills gaps · target groups · sectors · geographies and trends — feeding the Wizard and the portal.

03 USER ROLES

Policy makers
EU · national · regional

Researchers
labour economists · evaluators

PES & experts
trainers · counsellors

Pilot beneficiaries
trainees · job seekers

OUTPUTS

ALMP design documents
Skill-gap & trend reports
Evidence-backed answers

The Online Job Advertisement/ (OJA) processing pipeline

47 portals · 10+ countries · ~2–3 M postings · 11 stages

THE PROBLEM

Non-uniform free-text

Upstream stages extract structured fields using LLMs and translate to English — but terms are inconsistent: **"Excel"**, **"MS Office"**, **"spreadsheet skills"** all refer to the same concept.

OUR CONTRIBUTION

Stages 7 + 8 — normalisation

Attach **canonical taxonomy codes** to every extracted field via the normalisation service. Translation at stage 5 enables **cross-country comparability** — model-agnostic per language.

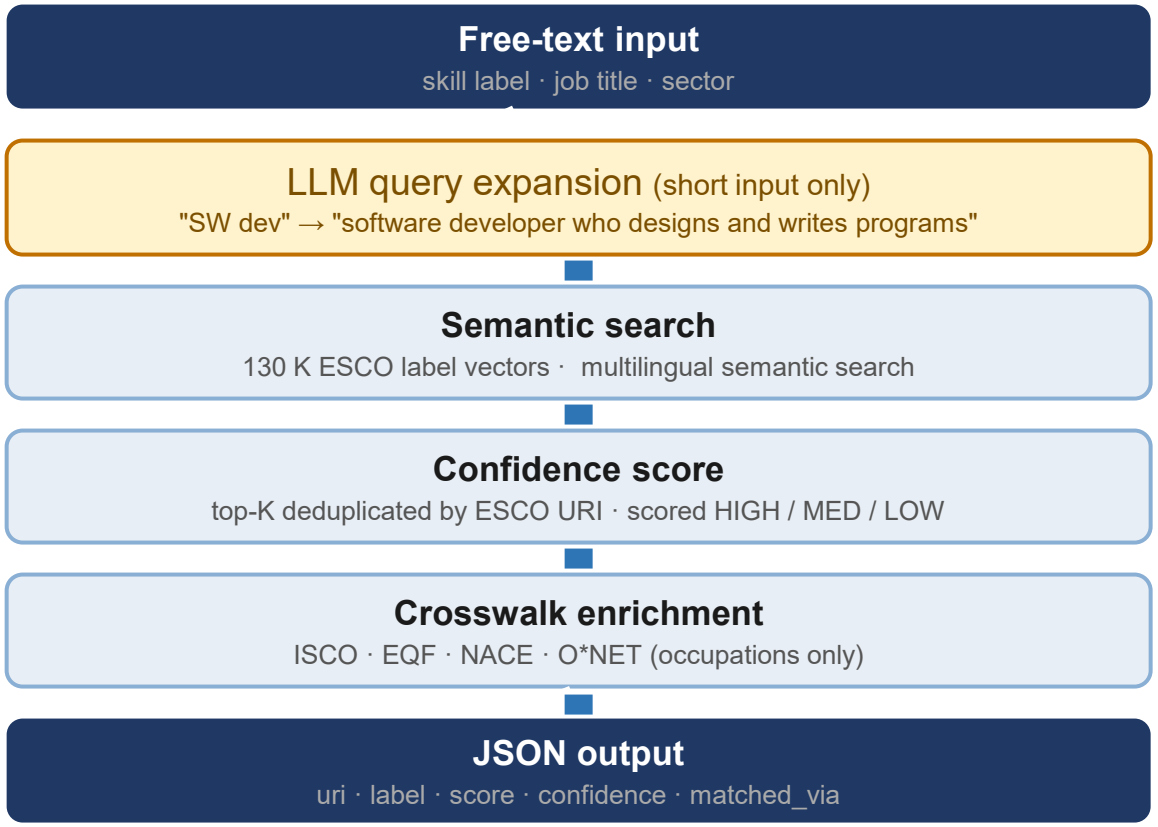
OUTPUT SCHEMA

Each field mapped to its taxonomy

Occupation	ESCO → ISCO-08
Skills	ESCO skills
Sector	NACE Rev.2.1
Education	EQF level
Digital / Green flags	DigComp · GreenComp

Normalization backend — Lexicon lookup and LLMs

Free-text input → semantic search → confidence routing → enriched JSON output



LLM #1 — QUERY EXPANSION

Fires on short or abbreviated input

"SW dev" → expands to full description before embedding. The expansion is what gets embedded — not the original query.

LLM #2 — CONFIDENCE ROUTING

HIGH ≥ 0.80 → accepted directly

Vector match is strong — result returned immediately, no LLM call.

MEDIUM 0.65 – 0.79 → accepted, flagged

Passes through with confidence marker. Consumer decides whether to review.

LOW < 0.65 → LLM judge fires

Receives top-5 **distinct ESCO URIs** (deduplicated due to prior overfetching). Picks the best match or rejects. Cannot select outside this set — grounded in ESCO, no hallucination.

Datasets used

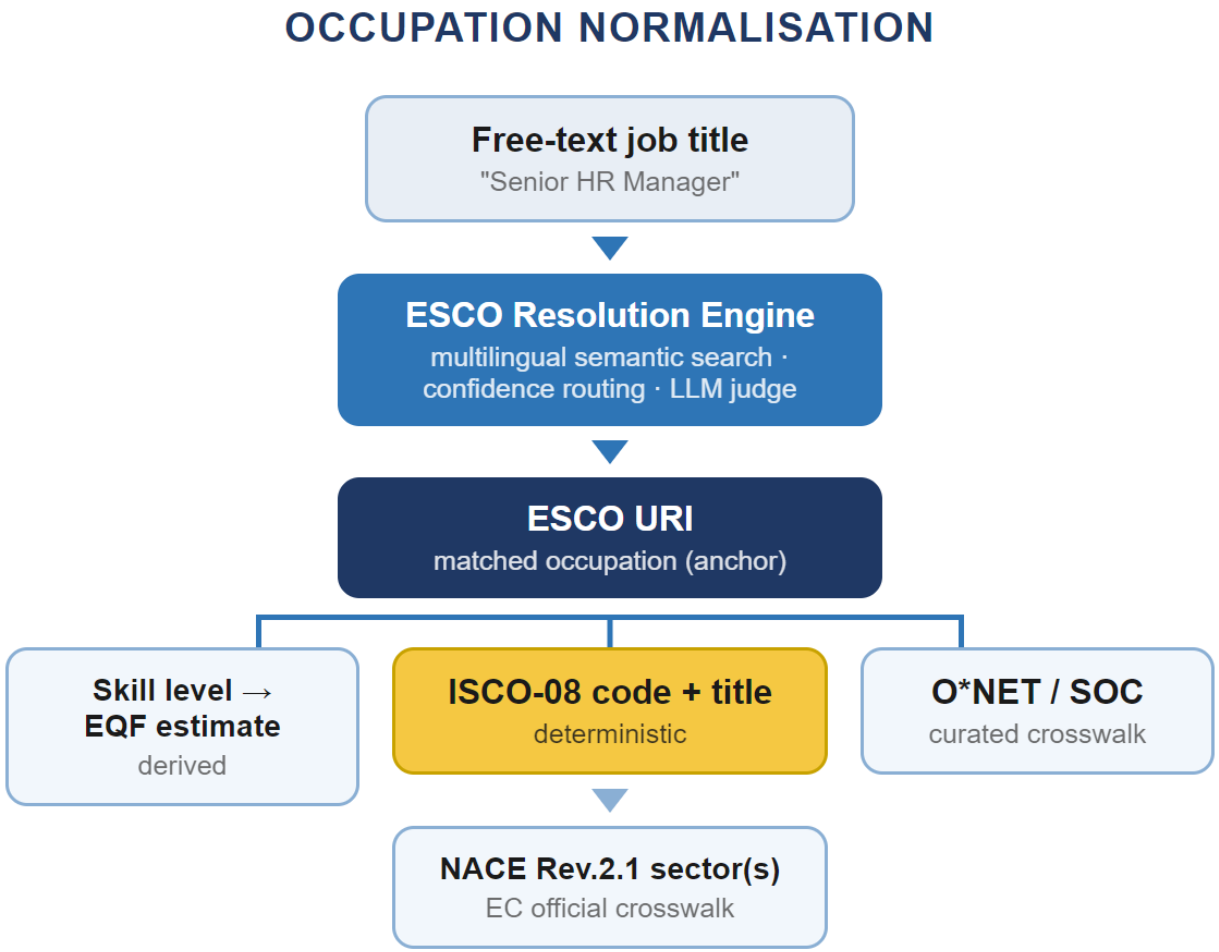
Dataset	Coverage	Role
ESCO v1.2.1 — occupations	3,043 entries	Anchor for occupation field
ESCO v1.2.1 — skills	13,960 entries	Anchor for skills field
ILO ISCO-08 group titles	619 groups (4 levels)	Canonical statistical titles
ESCO ↔ NACE Rev.2.1	4,632 mappings	Sector classification
GreenComp — green skills	630 entries	Marks skills as green
DigComp 2.2 — digital	1,285 entries	Marks skills as digital
Lightcast tech-skill lexicon	257 entries	Covers tech-skill gaps in ESCO

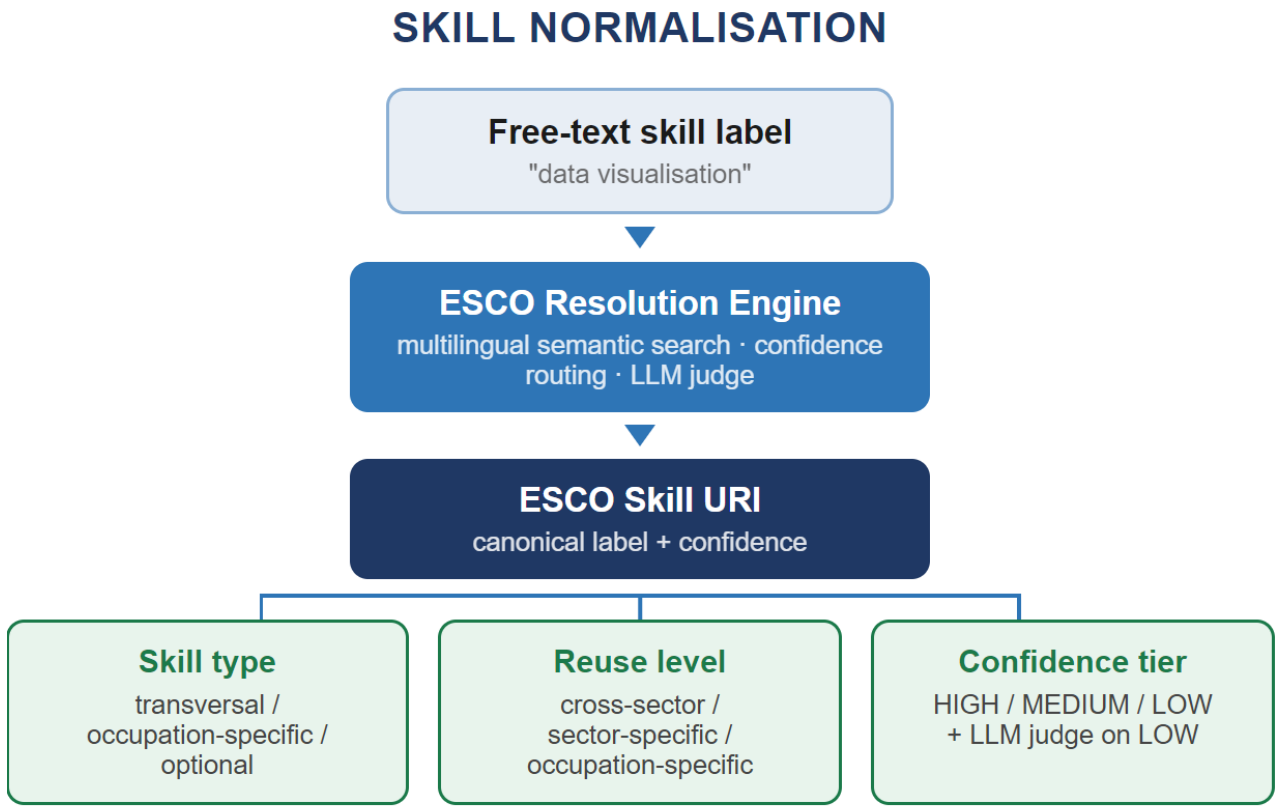
All datasets open license · CC BY 4.0 / EU Open Data · fully reproducible pipeline

Architecture overview

"Dual Normalisation Pipelines"

Occupations and skills resolved to canonical ESCO URIs via the same semantic search architectures





Evaluation — Purpose-Built Golden Test Suite

225 cases · 45 ESCO occupations · Title-only input — the hardest evaluation setting · 100% verified ground truth

OVERALL ACCURACY

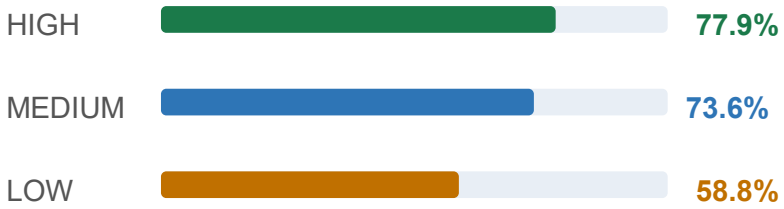
73.8%
ISCO-3 group correct
Eurostat / Cedefop reporting standard

55.6%
Exact ESCO URI match
Strict — same occupation, same URI

ISCO-3 ACCURACY BY TITLE DIFFICULTY

Difficulty	*Example*	ISCO-3
Easy	*"Software Developer"*	**86.7%**
Medium	*"Sr. PM - IT Dept"*	**80.0%**
Hard	*"Sr. SW Eng TL"*	**61.1%**

ISCO-3 ACCURACY BY CONFIDENCE TIER (*)



COVERAGE SUMMARY

92%
of OJAs receive an ESCO classification
HIGH + MEDIUM tiers included in statistics

HOW THE SYSTEM RESOLVES REAL TITLES

VECTOR MATCH — score ≥ 0.85, no LLM

`"Software Engineer"` → software developer **FAISS 0.91**
"engineer" is ESCO altLabel for "developer" — same occupation

`"HR Manager"` → human resources manager **FAISS 0.95**
altLabel hit — abbreviation in index

LLM JUDGE — fired on low score

`"Sr. HR Mgr"` → human resources manager **LLM 0.73**
LLM decoded abbreviation, picked from top-5

`"Ops Manager"` → operations manager **LLM 0.85**
LLM confirmed top candidate

HARD — heavy abbreviation, LLM decoded

`"Constr. Gen. Supv."` → construction general supervisor **LLM 0.54**
Low confidence — LLM expanded abbreviations

`"Data Anlst (BI)"` → data analyst **LLM 0.51**
Low confidence score low — LLM decoded & confirmed

(*) Title difficulty and prediction confidence are correlated but not identical; confidence is assigned by the model per prediction and does not directly correspond to difficulty level.

Skill Normalisation — Evaluation Results

Three-tier matching: vector retrieval → confidence filter → LLM judge where needed

78.3%
Semantic accuracy
LLM judge — equivalent concept

67%
Cases triggered LLM judge
avg latency 780ms

SKILLS VS OCCUPATIONS — SAME PIPELINE

Difficulty defined per domain: title complexity for occupations, label formality for skills.

Difficulty	*Example*	Occupation	Skill
Easy	*"Software Developer", "SQL"*	86.7%	91.7%
Medium	*"Sr. PM", "ML", "CRM"*	80.0%	87.5%
Hard	*"Sr. SW Eng TL", "Py/SQL"*	61.1%	62.5%
Overall		73.8%	78.3%

API RESPONSE INCLUDES LLM REASONING

```
"confidence": "medium",
"llm_judge": {

"triggered": true,

"reasoning": "PM² is the EU project management framework — maps to ESCO skill",

"match_found": true
}
```

HOW THE SYSTEM RESOLVES SKILL LABELS

DIRECT MATCH — canonical or altLabel

`"SQL"` → SQL **Vector 1.00**
exact preferred label hit

`"ML"` → machine learning **Vector 0.91**
altLabel "ML" in ESCO index

LLM JUDGE — abbreviation or informal term

`"CRM software"` → CRM / customer rel. mgmt. **LLM 0.78**
LLM expanded abbreviation, picked from top-5

`"PM² methodology"` → project management **LLM 0.71**
LLM resolved EU framework name to ESCO skill

HARD — synonym drift, near-miss

`"analytical mindset"` → critical thinking **LLM 0.52**
semantically close — different URI, counted as near-miss

`"Py/SQL scripting"` → SQL **LLM 0.61**
composite skill — system picks one, misses Python URI

Takeaways

1 **73.8% ISCO-3 accuracy · 92% of OJAs classified**

Semantic search handles most titles locally in ~2ms — LLM fires only where reasoning adds value, keeping EU-scale processing practical.

2 **One framework, four outputs in one pass**

Occupation (ESCO / ISCO) · Skills · Industry (NACE) · Education (EQF) — same pipeline, no reprocessing, one API call.

3 **Evaluation is a lower bound**

Tested on title-only input. Passing full OJA context — employer industry, listed skills — to the LLM judge is the next step, expected to push accuracy further.

Thank you.

CONTACT

Assoc. Prof. P. Zervas — University of Peloponnese
Email: p.zervas@uop.gr,
Deputy Scientific Coordinator
EU-ALMPO consortium

PROJECT

eu-almpo.eu

FUNDING

European Union · Horizon Europe
Grant Agreement Nº 101178736