The Journal of Open Source Software

LAiSER: A Taxonomy-Aware Framework for Skill Extraction and Research

Satya Phanindra Kumar Kalaga [1] **and Bharat Khandelwal** [1]

1 Program on Skills, Credentials and Workforce Policy, Institute of Public Policy, The George Washington University, USA

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 01 November 2025
Published: unpublished

Summary

LAiSER is an artificial intelligence framework for extracting, standardizing, and aligning skill information from unstructured text with established skill taxonomies. The system addresses the lack of standardized, machine-readable skill information needed to facilitate communication between learners, educators, and employers. LAiSER employs a two-stage pipeline combining large language models (LLMs) with semantic vector search using FAISS (Johnson et al., 2019) for high-precision skill extraction and taxonomy alignment to standards like ESCO (European Commission, 2020). The framework supports multiple computational backends—GPU-accelerated inference, cloud APIs, and CPU-only environments—making it accessible across diverse research and operational contexts.

Statement of need

The contemporary labor market is characterized by rapid technological change, evolving skill requirements, and growing disconnect between educational curricula and industry needs (Autor & Dorn, 2013). Traditional skill extraction approaches rely on manual annotation, expert judgment, or keyword matching—methods that are labor-intensive, error-prone, and fail to capture semantic richness (Boselli et al., 2018). LAiSER fills this gap by providing an accessible, flexible, and accurate framework for automated skill extraction and taxonomy alignment, enabling researchers, educators, and workforce professionals to analyze skill demands at scale.

Key Features

LAiSER provides three core innovations:

1. **Intelligent Text Preprocessing**: Domain-aware preprocessing removes extraneous information (company branding, legal boilerplate, benefits) while preserving task-relevant skill content.
2. **Multi-Model Skill Extraction**: Flexible architecture supporting multiple LLMs (vLLM, HuggingFace Transformers, Gemini API) with automatic fallback mechanisms for deployment across different computational environments.
3. **Taxonomy-Aware Alignment**: Semantic similarity-based alignment using sentence transformers (Reimers & Gurevych, 2019) and FAISS vector search maps extracted skills to standardized taxonomies, enabling cross-domain skill analysis.

The framework also supports extraction within the Knowledge, Skills, and Abilities (KSA) framework with automatic proficiency level classification using the Scottish Credit and Qualifications

Framework (SCQF) (Scottish Credit and Qualifications Framework Partnership, 2019).

Implementation

LAiSER is implemented in Python 3.9+ using PyTorch, HuggingFace Transformers, FAISS, spaCy, Sentence-Transformers, and pandas. The modular service architecture separates data access, skill extraction, and taxonomy alignment into loosely coupled layers.

The system accepts pandas DataFrames containing textual descriptions and produces structured output including raw extracted skills, taxonomy-aligned canonical skills, taxonomy identifiers (e.g., ESCO codes), and semantic similarity scores. For KSA extraction, output includes SCQF proficiency levels, knowledge requirements, and task abilities.

Applications

LAiSER enables diverse applications: labor market intelligence through large-scale job advertisement analysis; curriculum development through skill gap identification between educational programs and industry demands; skills-based hiring through standardized job-candidate matching; career pathway analysis through skill similarity assessment across occupations; and workforce policy research through regional skills assessments.

Availability

The source code is available at https://github.com/LAiSER-Software/extract-module under the MIT License. Installation via pip:

```
pip install laiser[gpu]  # GPU-accelerated
pip install laiser[cpu]  # CPU-only
```

Example usage:

```
from laiser.skill_extractor_refactored import SkillExtractorRefactored
import pandas as pd

extractor = SkillExtractorRefactored(model_id="gemini", api_key="YOUR_API_KEY", use_gpu=
data = pd.read_csv("job_descriptions.csv")
results = extractor.extract_and_align(data, id_column="job_id",
    text_columns=["description"], input_type="job_desc")
```

Acknowledgments

The authors acknowledge the George Washington University Institute of Public Policy and the Program on Skills, Credentials, and Workforce Policy for institutional support. This project was supported by grants from the Walmart Foundation and the Gates Foundation. The authors thank the GW Open Source Program Office and the developers of HuggingFace Transformers, FAISS, and spaCy.

References

Autor, D. H., & Dorn, D. (2013). The growth of low-skill service jobs and the polarization of the US labor market. *American Economic Review*, *103*(5), 1553–1597. https://doi.org/10.1257/aer.103.5.1553

Boselli, R., Cesarini, M., Mercorio, F., & Mezzanzanica, M. (2018). Classifying online job advertisements through machine learning. *Future Generation Computer Systems*, *86*, 319–328. https://doi.org/10.1016/j.future.2018.03.035

European Commission. (2020). *ESCO: European skills, competences, qualifications and occupations*. https://ec.europa.eu/esco

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, *7*(3), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. https://doi.org/10.18653/v1/D19-1410

Scottish Credit and Qualifications Framework Partnership. (2019). *SCQF handbook: User guide*. https://scqf.org.uk