# A Scalable Pipeline for Novelty Detection in Skill Extraction Using Large Language Models

**Gian Seifert, Simon Clematide**
University of Zurich, University of Zurich
Rämistrasse 71, 8006 Zürich, Andreasstrasse 15, 8050 Zurich, Switzerland
gian.seifert@uzh.ch, simon.clematide@cl.uzh.ch

**Abstract**

The rapid evolution of the labor market requires skill ontologies to be continuously updated, but manually identifying emerging skills in job advertisements is highly labor-intensive. This paper presents a scalable, multi-stage pipeline for automated novelty detection in skill extraction. The system combines Large Language Models (LLMs) for candidate generation, a re-matching and threshold-based filtering module ("Turbo"), that compares candidates against the existing ontology, and a two-step aggregation process that merges string-based and embedding-based clustering. Experiments on Swiss job advertisement datasets using GPT-4o, Gemini-2.0-flash, and DeepSeek-V3 show that the pipeline effectively reduces noise and manual curation effort: Turbo filtering lowered false positives by 82%, and aggregation reduced the number of items requiring review by 97%. Among the tested models, Gemini-2.0-flash achieved the highest precision, reaching a novelty detection ratio of up to 73% in the qualitative evaluation. These findings demonstrate the pipeline's potential as an efficient tool for maintaining dynamic skill ontologies.

**Keywords:** Novelty Detection, Skill Extraction, Large Language Models, Ontology Maintenance, Natural Language Processing

## 1. Introduction

The shift towards skill-based hiring demands comprehensive and up-to-date skill ontologies. However, the sheer volume and rapid evolution of job advertisements make manual ontology maintenance a significant challenge. This paper addresses this issue by proposing an automated pipeline to detect skills that are not yet represented in an existing ontology, ensuring its continued relevance. Recent advancements in Large Language Models (LLMs) offer a promising approach to this problem. Their strong ability to understand the semantic context of natural language makes them well-suited for identifying novel skills that do not fit existing terminologies or patterns. We investigate how LLMs can support the detection of novel skills and thereby bridge automated extraction and expert curation.

## 2. Related Work

Skill extraction is a research area within Natural Language Processing (NLP) that focuses on identifying skills from textual sources such as job postings (Zhang et al., 2022). The adoption of structured skill ontologies, such as the multilingual ESCO (European Skills, Competences, Qualifications and Occupations) in Europe and O*NET (Occupational Information Network) in the United States, has considerably improved both the precision and standardization of this task (Commission et al., 2017; Gregory et al., 2019). Novelty detection, the task of identifying data points that deviate from established patterns, is an important topic in machine learning (Pimentel et al., 2014). Existing approaches range from statistical and traditional machine learning methods to recent deep learning architectures. The advent of LLMs has further advanced novelty detection in NLP by enabling a more nuanced understanding of context and semantics (Ghosal et al., 2022). However, applying LLMs specifically to novelty detection within skill extraction pipelines remains a relatively new line of research. Furthermore, this work relates to the broader field of Ontology Matching (OM). While traditional OM relies on lexical and structural alignments, recent work has explored the use of LLMs to improve semantic matching capabilities (Taboada et al., 2025). The continued relevance of this area is reflected in the Ontology Alignment Evaluation Initiative (OAEI) 2025 campaigns. Our pipeline contributes to this line of research by framing novelty detection as an extension of ontology matching, identifying concepts that fall outside existing alignments.

## 3. Data and Methods

### 3.1. Data

Three datasets were derived from a collection of more than 20 million Swiss job ads for use in the experiments:

**10_isco** A small, diverse sample of 10 German-language job ads, each representing a distinct major ISCO-08 occupational group This dataset was created for a broad overview of all occupational groups. Despite its limited size, this sample was



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 7701–7706
11–16 May 2026. 

highly curated, as the substantial cost of manual annotation required a smaller dataset to maintain review quality.

**30_industry_after_2024** A multilingual set of 30 job ads from sectors with high expected innovation (e.g., ICT, medical technology), all published after January 1, 2024. We created this dataset to maximize the likelihood of finding novelties. As with the ISCO dataset, this sample size was appropriate for an exploratory study intended to assess the pipeline's potential.

**1000_industry_after_2024** An extended version of the previous dataset containing 1000 job ads with the same selection criteria, used for tuning and testing the aggregation step.

All datasets originate from a proprietary corpus of over twenty million Swiss job advertisements collected since 2012. The corpus is available for research under license but is not publicly accessible. However, to promote reproducibility, a subset of the data, our prompt templates, and the complete pipeline code will be released via a public GitHub repository[1].

### 3.2. Existing Skill Extraction Pipeline

The proposed novelty detection system extends an existing three-step skill extraction pipeline:

**Zoning**: A BERT-based model identifies text zones relevant to skills and experience within job advertisements.

**Skill Mention Extraction**: An LLM extracts skill mentions and formats them according to the topic-predicate schema used in the ontology.

**Ontology Matching**: A text-embedding-based algorithm maps the extracted skill mentions to the existing skills ontology, producing a ranked list of ontology-matched skills. The algorithm assigns each pair a similarity distance between 0 and 8, with lower values indicating greater similarity. German example pairs for this distance metric are shown below:

*"Prophylaxe Lektionen durchführen"* ("Conduct prophylaxis lessons") vs. *"Prophylaxe durchführen"* ("Carry out prophylaxis") – distance 0.6;
*"Offene Haltung bewahren"* ("Maintain an open attitude") vs. *"Ansehen bewahren"* ("Maintain reputation") – distance 1.4;
*"Teamspirit teilen"* ("Share team spirit") vs. *"Teams formen"* ("Form teams") – distance 1.7.

### 3.3. Proposed Novelty Detection Pipeline

The proposed pipeline for detecting novel skills comprises three main stages:

[1] https://github.com/gseife/Novelty-Detection-in-Skill-Extraction

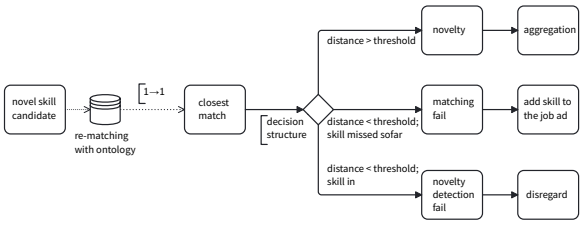


Figure 1: Data flow for *Turbo* processing steps and decision structure

**Candidate Generation** An LLM is prompted to compare the list of extracted skill mentions with the list of ontology-matched skills, using the full job ad text as context. Its task is to identify mismatches and generate *novel skill candidates*—skills that appear in the text but are absent from the ontology matches.

The quality of this stage depends strongly on careful prompt engineering. A multi-component prompt was developed to guide the LLM's behavior and ensure that its outputs are both relevant and structured for downstream processing. The main elements of the prompt design are:

**Persona Setting:** The system prompt instructs the LLM to act as an "expert in the Swiss labor market and in skill extraction." This grounds the model's analysis in a defined professional and geographical context, improving the relevance of its outputs.

**Structured Input:** The prompt provides the lists of ontology-matched skills and skill mentions, together with the full job advertisement, job title, and timestamp. Placing the advertisement text at the end of the prompt produced more consistent results, as it allows the model to process the task instructions before reading the contextual information.

**Output Formatting:** To maintain compatibility with the target ontology and enable automated parsing, the prompt enforces a strict output format, prioritizing ease of downstream processing at the cost of a slight performance trade-off (Tam et al., 2024). Novel skills must follow the topic-predicate schema [e.g., Original:"Python programmieren" (translation:"Python programming")] and be returned as a JSON object.

**Turbo: Post-Processing and Filtering** This module serves as an effective filter that refines the pool of candidate skills. Each candidate is re-matched against the complete skills ontology using an embedding-based algorithm that computes a similarity distance. A configurable threshold is then applied to determine whether a candidate is retained.

The overall data flow and decision structure are

illustrated in Figure 1. The figure provides an overview of how candidates are processed and filtered, while the following points summarize the decision logic:

- **Above Threshold (Novelty):** The candidate has no close match in the ontology and is classified as a genuine novelty. Example: *Kernelemente konzipieren* (translation:"Designing core elements").

- **Below Threshold (Not a Novelty):** The candidate is similar to an existing skill and is further divided into:

  - **Detection Fail** The re-matched skill already appears in the initial list of ontology-matched skills for that job ad, indicating a false positive from the LLM. Example: *IT-Prozesse automatisieren* (translation:"Automate IT processes")

  - **Matching Fail** The re-matched skill exists in the ontology but was missed by the initial matching step, revealing a limitation in the upstream process rather than a new skill. Example: *IT-Lösungen integrieren* (translation:"Integrate IT solutions").

**Aggregation of Novelties** To minimize the remaining manual curation effort, all skills classified as novelties are aggregated through a two-step process:

**Token Sort Aggregation:** Lexically and syntactically similar skills are grouped using the FuzzyWuzzy library's token sort algorithm (Cohen, 2011), which handles minor variations in word order (e.g., "Data Science Expert" vs. "Expert in Data Science").

**Embedding-based Clustering:** Semantically related but infrequent skills (the long tail) are grouped using DBSCAN (density-based spatial clustering) applied to their text embeddings (OpenAI's text-embedding-3-small) (OpenAI, 2024; Ester et al., 1996). This step merges skills with similar meanings but different surface forms.

## 4. Results and Discussion

The proposed pipeline was evaluated using three different LLMs: GPT-4o (OpenAI et al., 2024), DeepSeek-V3 (DeepSeek-AI et al., 2025), and Gemini-2.0-flash (Google, 2025). Each model was tested on Swiss job advertisement datasets to evaluate its ability to detect novel skills and to assess the effectiveness of the filtering and aggregation components.

### 4.1. Quantitative Pipeline Evaluation

We assess performance in terms of three factors: (1) the number of novel skill candidates generated, (2) the filtering capacity of the *Turbo* module, and (3) the reduction in manual curation through aggregation.

**Candidate Generation** As an initial analysis, we examined how many novel skill candidates each LLM produced prior to filtering. On the 30_industry_after_2024 dataset, DeepSeek-V3 generated 233 candidates, Gemini-2.0-flash 131, and GPT-4o 77. These results show clear differences in model behavior and highlight that a large proportion of suggestions were noisy or irrelevant, reinforcing the need for a dedicated post-processing component such as the *Turbo*.

***Turbo*: Post-Processing and Filtering** A separate evaluation was conducted to assess the *Turbo* component as a post-processing mechanism within the novelty detection pipeline. Its function is to reduce false positives while preserving the detection of genuine novelties. Three similarity threshold levels were tested: strict (1.7), moderate (1.4), and lenient (1.3), to identify an optimal balance between manual review effort and recall. Thresholds were selected based on the distribution of distances between novel skill candidates and their closest ontology matches. A threshold of 1.7 corresponds to approximately the top 5% most distant candidates, while thresholds of 1.4 and 1.3 correspond to approximately 15% and 20%, respectively. These percentile-based cutoffs were chosen after a preliminary empirical assessment of the results and correspond to increasingly permissive settings.

The quantitative analysis (Table 1) shows strong filtering performance, expressed as the inverse of the novelty ratio (defined here as the proportion of generated candidates that were retained by Turbo as novelties, not yet as manually validated true novelties). At the strictest threshold of 1.7, the *Turbo* filtered out 97% of GPT-4o candidates, 93% of Gemini-2.0-flash, and 84% of DeepSeek-V3, indicating substantial noise reduction.

A key evaluation metric is the Detection Fail Ratio, which quantifies the proportion of false positives (defined as the percentage of re-matched candidates that already appeared in the initial ontology matches for the same job ad). Gemini-2.0-flash consistently achieved the lowest false positive rate across all thresholds (e.g., 33% at 1.4), demonstrating the highest precision. In contrast, DeepSeek-V3 produced the highest rate (60% at 1.4), confirming its tendency to over-generate candidates.

The configurable threshold of the *Turbo* enables a controlled trade-off between precision and recall,

| Threshold | Model | ↑ Novelty | ↑ Match Fail | ↓ Detection Fail |
|---|---|---|---|---|
| 1.7 | GPT-4o | 3 | 43 | 55 |
| | DeepSeek-V3 | 16 | 7 | 76 |
| | Gemini-2.0-flash | 7 | 57 | 36 |
| 1.4 | GPT-4o | 16 | 39 | 45 |
| | DeepSeek-V3 | 34 | 6 | 60 |
| | Gemini-2.0-flash | 18 | 49 | 33 |
| 1.3 | GPT-4o | 23 | 36 | 40 |
| | DeepSeek-V3 | 39 | 6 | 54 |
| | Gemini-2.0-flash | 27 | 44 | 29 |

Table 1: Quantitative comparison on the **30_industry_after_2024** dataset. Values show the novelty, matching-fail, and detection-fail rates in percent for each model and threshold.

| Threshold | Model | Novelty Ratio ↑ | Correct-Format Ratio ↑ |
|---|---|---|---|
| 1.4 | DeepSeek-V3 | 27 | 4 |
| | Gemini-2.0-flash | **73** | 45 |
| | GPT-4o | 67 | **67** |
| 1.3 | Gemini-2.0-flash | **72** | 44 |
| | GPT-4o | 60 | **67** |
| 1.2 | Gemini-2.0-flash | **62** | 46 |
| | GPT-4o | 57 | **61** |

Table 2: Qualitative model comparison across thresholds 1.4, 1.3, and 1.2 on the **30_industry_after_2024** dataset. Values show novelty and correct-format rates in percent for each model and threshold.

allowing adaptation to different operational contexts. Higher thresholds are suitable for high-precision production settings, whereas lower thresholds are appropriate for exploratory phases with more extensive human validation. Although strict thresholds may exclude some genuine novelties, this effect is mitigated by the redundancy of job advertisement data, as true skills typically recur across multiple postings.

**Aggregation of Novelties** The final aggregation step was introduced to minimize the number of skills requiring manual expert review. When evaluated on the larger 1000_industry_after_2024 dataset, it achieved a 97% reduction in the total number of individual skills that needed curation. After aggregation, 12% of the initially proposed novel skills were retained within the resulting clusters. These results demonstrate that the aggregation process effectively streamlines manual review by consolidating recurring novelties while preserving the essential information discovered by the pipeline.

### 4.2. Qualitative Evaluation

**Evaluation Criteria** To evaluate the quality of the generated novelties, all outputs were manually reviewed and classified. Given the pilot nature of this evaluation, human validation was conducted as an exploratory manual review rather than according to a strict multi-annotator protocol. All candidate items generated across the models were reviewed by the first author, with ambiguous cases discussed in consultation with a skills expert. The analysis revealed four main categories of matches:

**Match Fail:** A suitable equivalent already existed in the ontology but was not retrieved during the initial matching step.

**Multiple Skills in Skill Mention:** The extracted mention contained several distinct skills, but the one-to-one matching algorithm captured only one, causing the LLM to flag the remaining ones.

**Missing Skill-Predicate Combination (True Novelty):** The topic and predicate were both present in the ontology but not in the specific combination required.

**Topic Fail (True Novelty):** The main concept or topic of the skill was entirely absent from the ontology.

**Model Comparison** The comparative evaluation of LLMs for true-novelty detection revealed consistent differences in precision and reliability across *Turbo* threshold levels. Gemini-2.0-flash achieved the highest novelty detection ratios, 73%, 72%, and 62% at thresholds 1.4, 1.3, and 1.2 respectively, demonstrating strong capability in identifying genuine novelties.

GPT-4o produced the most consistently formatted skill candidates, achieving a correct-format ratio (defined as the percentage of outputs that strictly adhered to the requested topic-predicate schema) of 67% at thresholds 1.4 and 1.3, and 61% at 1.2. Its novelty detection performance, however, was moderate. DeepSeek-V3 generated the largest volume of candidates but exhibited the lowest precision, with a 27% novelty ratio, a 4% correct-format ratio, and the highest false-positive rate (up to 76% at threshold 1.7). False-positive analysis confirmed Gemini's superior precision and DeepSeek's tendency to over-generate.

Overall, these results indicate a trade-off between structural accuracy and novelty detection. GPT-4o excels in producing syntactically correct outputs, whereas Gemini-2.0-flash performs best in identifying true novelties with only a minor loss in output structure quality. DeepSeek-V3, while prolific in candidate generation, lacks the precision required for practical application. Based on these findings, Gemini-2.0-flash represents the most effective model for high-precision novelty detection in this context.

## 5. Conclusion

This study has presented and validated a scalable, multi-stage pipeline for novelty detection in skill extraction using LLMs. The approach balances automation with expert oversight, demonstrating that

LLMs, when integrated into a structured framework of filtering and aggregation, can serve as an effective tool for maintaining dynamic skill ontologies.

The main contributions of this work are threefold:

1. LLMs can detect novel skills effectively, with the best-performing model achieving a novelty detection ratio of up to 73%.
2. A robust filtering and aggregation methodology reduces the manual workload of ontology curation by 97%.
3. A comparative evaluation of several LLMs provides practical guidance on model selection and assessment for this task.

Despite these results, several limitations remain. The evaluation was conducted on a specific set of LLMs at a single point in time; given the rapid development of this technology, performance benchmarks will evolve quickly. Additionally, the pipeline relies heavily on prompt engineering and is therefore sensitive to prompt design. However, the evaluation framework established for this dataset should make it easier to refine prompts and adapt the pipeline to newer LLMs. Furthermore, the current study is limited to Swiss job advertisements and to the task of skill detection. In addition, the aggregation step requires dataset-specific parameter tuning, which currently depends on manual configuration. An additional expert review also highlighted several directions for future work, particularly regarding integration into curation tools, the continued role of human oversight, and the distinction between ontology blind spots and genuinely emerging skills.

These limitations, together with the evaluation insights, point to promising directions for future research. These include (1) automated parameter tuning to adapt aggregation settings over time, (2) applying the pipeline to other standard skill ontologies such as ESCO or O*NET, (3) exploring hybrid model setups where different LLMs handle subtasks according to their strengths, and (4) validating the pipeline's generalizability across non-Swiss datasets, other languages, and different domains, such as medical or scientific texts.

This work establishes a strong foundation for applying LLMs in automated novelty detection and continuous ontology maintenance. The proposed pipeline, through its adaptable architecture and reliable performance, offers a practical framework for keeping skill ontologies dynamic, relevant, and aligned with the evolving demands of the labor market.

## 6. Acknowledgments

We thank the team at x28 AG, especially Dr. Felix Busch and Flavio Battaini, for their collaboration and support in this work.

## 7. References

Adam Cohen. 2011. Fuzzy String Matching in Python. Accessed: 2025-05-21.

European Commission, Social Affairs Directorate-General for Employment, and Inclusion. 2017. *ESCO handbook – European skills, competences, qualifications and occupations*. Publications Office.

DeepSeek-AI, Aixin Liu, Bei Feng, et al. 2025. DeepSeek-V3 Technical Report.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press.

Tirthankar Ghosal, Tanik Saikh, Tameesh Biswas, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Novelty Detection: A Perspective from Natural Language Processing. *Computational Linguistics*, 48(1):77–117. _eprint: https://direct.mit.edu/coli/article-pdf/48/1/77/2006641/coli_a_00429.pdf.

Google. 2025. Gemini 2.0 Flash. Accessed: 2025-06-19.

Christina Gregory, Phil Lewis, Pamela Frugoli, and Alexander Nallin. 2019. Updating the O*NET®-SOC Taxonomy: Incorporating the 2018 SOC Structure. Technical report, National Center for O*NET Development, Raleigh, NC. Accessed: 2025-06-06.

OpenAI. 2024. text-embedding-3-small. Accessed: 2025-05-21.

OpenAI, Aaron Hurst, Adam Lerer, , et al. 2024. GPT-4o System Card.

Marco A. F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. 2014. A review of novelty detection. *Signal Processing*, 99:215–249.

Maria Taboada, Diego Martinez, Mohammed Arideh, and Rosa Mosquera. 2025. Ontology matching with Large Language Models and prioritized depth-first search. *Information Fusion*, 123:103254.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung yi Lee, and Yun-Nung Chen. 2024. Let Me Speak Freely? A Study on the Impact of Format Restrictions on Performance of Large Language Models.

Mike Zhang, Kristian Nørgaard Jensen, Rob van der Goot, and Barbara Plank. 2022. Skill Extraction from Job Postings using Weak Supervision. ArXiv:2209.08071 [cs].